ACTA UNIVERSITATIS UPSALIENSIS
Studia Linguistica Upsaliensia
15

Discourse in Statistical Machine Translation

Christian Hardmeier

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in Universitetshuset, Sal X, Uppsala, Saturday, 14 June 2014 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.
Faculty examiner: Dr. Lluís Màrquez (Qatar Computing Research Institute).

Abstract
Hardmeier, C. 2014. Discourse in Statistical Machine Translation. *Studia Linguistica Upsaliensia* 15. 185 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-554-8963-2.

This thesis addresses the technical and linguistic aspects of discourse-level processing in phrase-based statistical machine translation (SMT). Connected texts can have complex text-level linguistic dependencies across sentences that must be preserved in translation. However, the models and algorithms of SMT are pervaded by locality assumptions. In a standard SMT setup, no model has more complex dependencies than an *n*-gram model. The popular stack decoding algorithm exploits this fact to implement efficient search with a dynamic programming technique. This is a serious technical obstacle to discourse-level modelling in SMT.

From a technical viewpoint, the main contribution of our work is the development of a document-level decoder based on stochastic local search that translates a complete document as a single unit. The decoder starts with an initial translation of the document, created randomly or by running a stack decoder, and refines it with a sequence of elementary operations. After each step, the current translation is scored by a set of feature models with access to the full document context and its translation. We demonstrate the viability of this decoding approach for different document-level models.

From a linguistic viewpoint, we focus on the problem of translating pronominal anaphora. After investigating the properties and challenges of the pronoun translation task both theoretically and by studying corpus data, a neural network model for cross-lingual pronoun prediction is presented. This network jointly performs anaphora resolution and pronoun prediction and is trained on bilingual corpus data only, with no need for manual coreference annotations. The network is then integrated as a feature model in the document-level SMT decoder and tested in an English–French SMT system. We show that the pronoun prediction network model more adequately represents discourse-level dependencies for less frequent pronouns than a simpler maximum entropy baseline with separate coreference resolution.

By creating a framework for experimenting with discourse-level features in SMT, this work contributes to a long-term perspective that strives for more thorough modelling of complex linguistic phenomena in translation. Our results on pronoun translation shed new light on a challenging, but essential problem in machine translation that is as yet unsolved.

Keywords: Statistical machine translation, Discourse-level machine translation, Document decoding, Local search, Pronominal anaphora, Pronoun translation, Neural networks

Christian Hardmeier, Uppsala University, Department of Linguistics and Philology, Box 635, SE-75126 Uppsala, Sweden.

ISSN 1652-1366
ISBN 978-91-554-8963-2
urn:nbn:se:uu:diva-223798 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-223798)

Printed by Elanders Sverige AB, 2014

To Ursula

Contents

Acknowledgements

While working on this thesis, I received help and encouragement from many people. First and foremost, credit is due to my advisors. I began my Ph. D. studies at Fondazione Bruno Kessler (FBK) in Trento (Italy) in 2009 and completed them at Uppsala University after moving to Sweden in 2011. I had the great privilege of working with excellent and very dependable advisors in both places.

In Uppsala, *Joakim Nivre* and *Jörg Tiedemann* supported me with great involvement and care while allowing me complete freedom to pursue my own ideas. Having access to their combined competence and experience at every one of our meetings was an absolutely invaluable asset. Joakim taught me to unite visionary research goals with rigorous attention to detail, and Jörg constantly contributed new ideas to improve my methods and references to literature I did not know about. I benefited immensely from working with the two of them together.

During my two years in Trento, I enjoyed the supervision of *Marcello Federico*. He did his best to make me feel welcome in Trento and provided me with both equipment and opportunity to explore the skiing grounds on Monte Bondone. Careful and systematic, he would always be ready to discuss implementation details and raw experimental results or complex proofs in the derivation of statistical models. Much of what I know about the engineering aspects of statistical machine translation, I learnt from him.

Among my colleagues at work, two stand out particularly. In Uppsala, *Sara Stymne* discussed much of my work with me and freely contributed her advice. She acted as examiner at the mock defence preceding the submission of my thesis, proofread the entire manuscript and helped me address weaknesses in my results and their presentation. I am greatly indebted to her for her assistance in the final stages of preparing this thesis. In Trento, *Arianna Bisazza* was an excellent colleague and a good friend to me. I often missed her and our discussions on all kinds of linguistic and technical subjects after leaving Italy.

Two of my colleagues at Uppsala University's Department of Linguistics and Philology, *Marie Dubremetz* and *Mats Dahllöf*, annotated French pronouns for me. Marie also advised me on other matters requiring the linguistic competence of a native French speaker. Both my work and my social life in Uppsala became more interesting and enjoyable thanks to *Ali Basirat*, *Beáta Megyesi*, *Eva Martínez*, *Eva Petterson*, *Evelina Andersson*, *Marco Kuhlmann*,

Maryam Nourzaei, *Matthias Zumpe*, *Mattias Nilsson*, *Miguel Ballesteros*, *Mojgan Seraji*, *Oscar Täckström*, *Per Starbäck*, *Reut Tsarfaty*, *Sebastian Schleussner*, *Ute Bohnacker* and *Vera Wilhelmsen*.

During my time in Trento, *Nicola Bertoldi*, *Mauro Cettolo* and *Gabriele Musillo* of the FBK machine translation group had their share in discussions related to the early stages of my work. *Roldano Cattoni* was very helpful and patient with me when I used or abused the computing cluster at FBK.

My interest in statistical machine translation was first kindled by *Martin Volk*, who supervised my M. A. thesis on machine translation for film subtitles. He offered me much support even after I had taken up my Ph. D. studies in Trento, not least by repeatedly welcoming me as a summertime visitor at the University of Zurich and by contributing resources for the benefit of my research. During my stays in Zurich, I received much help with my experiments from *Rico Sennrich*, *Don Tuggener* and *Manfred Klenner*.

Soon after I published my first paper on pronouns in statistical machine translation, *Bonnie Webber* started taking a lively interest in my work and shared bits and pieces of her outstanding knowledge about all things related to discourse with me. Many times, her remarks made me gain a deeper understanding of the linguistic aspects of the phenomena I was dealing with.

My experiments used substantial computational resources. They were possible only because I had the opportunity to use two high-performance computing clusters in Oslo and Uppsala.[1] I am indebted to *Stephan Oepen* for permitting me to use a very generous part of his computing time allowance on the Abel cluster and to the system administrators of Abel for letting me overdraw my disk quota significantly while completing my thesis work.

When I arrived in Sweden in 2011, I was on my own and did not even have a place to stay, but luckily I had a faithful friend in Stockholm. I was warmly welcomed by *Roland Engkvist*, who offered me shelter in the maid's chamber of his flat on Kungsholmen until I could move to a more permanent place.

I am grateful to my parents, who inspired a scientific interest in me and supported my academic career in various ways throughout my life, and to my sister, to whom I owe much of what I know about translation studies and who proofread large parts of this thesis. Last but not least, my life would not be the same without *Ursula*, whose support and affection helped me through all these years. Thank you for being with me!

Uppsala, April 2014
Christian Hardmeier

[1] Computations were carried out on the Abel cluster, owned by the University of Oslo and the Norwegian metacenter for High Performance Computing (NOTUR) and operated by the Department for Research Computing at USIT, the University of Oslo IT department, under project nn9106k, as well as on resources provided by SNIC through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under project p2011020.

1. Introduction

Machine translation (MT) is the automatic translation of texts between natural languages by a computer system. Translation is a challenging task for humans, and it is no less challenging for computers. High-quality translation requires a thorough understanding of the source text and its intended function as well as good knowledge of the target language. In an MT system, this process must be completely formalised, which is a daunting task since the process is by no means completely understood. Statistical machine translation (SMT) addresses this challenge by analysing the output of human translators with statistical methods and extracting implicit information about the translation process from corpora of translated texts. SMT has shown good results for many language pairs and has had its share in the recent surge in popularity of MT among the general public.

Notwithstanding their success in practical translation scenarios, the methods used in SMT are shaped far more by technical constraints than by linguistic concerns. To ensure computational efficiency and tractability, complex linguistic interrelations are sacrificed to crude independence assumptions. The performance level of current SMT systems bears an amazing testimony to the fact that most information in natural languages is encoded very locally. Even though the context a typical SMT system considers is extremely impoverished, a great deal of information is usually transferred successfully into the target language. Nevertheless, human translators know that it is not sufficient to translate groups of words or sentences in isolation if a coherent target text is desired. In this thesis, we study some of the limitations of current SMT systems, in particular the implications of translating texts as sentences in isolation, as SMT systems usually do. We explore ways to overcome this limitation and investigate how cross-sentence, discourse-level context can be exploited in automatic translation.

1.1 Motivation and Goals

The point of departure of our research is the observation of a discrepancy between the fields of translation studies and machine translation. While it might seem that there should be strong connections between the two research areas, even a superficial look at the relevant literature quickly reveals that the two fields are preoccupied with completely different problems. In translation studies, much work has been devoted to defining and exploring the nature of

translation. It has been recognised since antiquity that word-by-word translation is generally inadequate and that a higher level of understanding is necessary to render a text adequately into another language. Confronted with the accusation of having taken liberties with the texts he translated from Greek into Latin, the fourth century church father and bible translator Jerome retorts:

> Ego enim non solum fateor, sed libera voce profiteor, me in interpretatione Graecorum, absque Scripturis sanctis, ubi et verborum ordo mysterium est, non verbum e verbo, sed sensum exprimere de sensu. (Jerome, 1996)
>
> For I myself not only admit but freely proclaim that in translating from the Greek (except in the case of the holy scriptures where even the order of the words is a mystery) I render sense for sense and not word for word. (Jerome, 1979)

Jerome defends his attitude by referring to the example of eminent writers of Roman antiquity like Cicero and Horace. His distinction between word-by-word and sense-by-sense translation was fundamental for theoretical discussions of translation until the first half of the 20th century (Bassnett, 2011).

The 20th century saw the rise of translation studies as a scientific discipline in its own right. Translation research began to focus on more precise and formal notions of translational equivalence such as the concept of dynamic equivalence advocated by Nida and Taber (1969), which seeks the object of equivalence at a pragmatic or functional level highly dependent on the message and intention of the source text and the reception of the target text. More recent theories of translation go even further and dispute the concept of equivalence altogether (Snell-Hornby, 1995), focusing instead on the cultural and social context and the intentionality of the production of both the original source text and the translation. The question of equivalence at the level of individual linguistic signs is an aspect of translator training (e. g., Baker, 2011, Chapter 2), but it does not meet with much interest otherwise; while good dictionaries are essential also for the human translator, their creation is largely the concern of lexicographers, not translation researchers.

The vast majority of the existing research on SMT, by contrast, is characterised by a happy disregard for the functional and pragmatic aspects of language. Instead, it deals with far more fundamental concerns such as the problem of generating grammatical word order in the target language. Much of the SMT research literature is fairly technically-minded and is concerned with finding more effective ways of applying existing statistical methods and techniques to the MT task without spending too much thought on the effects of using these methods on perceived translation quality.

Despite this discrepancy between how translation studies and SMT research approach the translation process, SMT has reached a point of maturity that enables it to be used by professional users in productive environments.

We suggest that it now makes sense for SMT researchers to take a step beyond what has been done traditionally and consider removing some of the restrictions that have been taken for granted in order to narrow the gap between SMT and the world of professional translators. One obvious step to take is the one from sentence-level translation to discourse. Most SMT research of the last twenty years has limited the context considered when generating a translation to that of the current sentence. While this restriction was adopted for sound technical reasons, it is a strong obstacle to the study of higher-level problems in SMT.

The standard models of SMT know very little about the linguistic structure of a text. Instead, when generating a part of their output, they exhaustively explore a context window around the current position, comparing translation variants and output word permutations and selecting the option that seems optimal given a set of models. To ensure tractability, the context window that is explored in this way must be kept small. In practice, SMT considers windows of no more than a handful of words. Once the context window has been reduced to this size, even more efficiency can be gained by using algorithms that specifically exploit the extreme locality of the context. This is a core feature of all commonly used decoding algorithms in SMT.

The primary goal of our research is to find ways around the sentence-level restriction in SMT and to explore how a larger context can be exploited to improve the quality of automatic translation. This problem has two aspects, both of which must be addressed to achieve an improvement in translation. If we wish to exploit unlimited discourse context in our SMT systems, we must *develop frameworks, procedures and algorithms that are not encumbered by the standard assumptions of sentence-level independence.* This is the first major research goal and the topic of the first part of this thesis. Our main contribution related to this goal is the development of a document-level decoding algorithm for phrase-based SMT. We have released software implementing this algorithm to the public in the form of our document-level phrase-based SMT decoder Docent (Hardmeier et al., 2013a) to provide a framework for the development of discourse-level SMT models for ourselves and other researchers.

With this essential piece of infrastructure in place, the next step is to *investigate what discourse-level linguistic phenomena can be useful for SMT, and how to model them in an SMT system.* We explore a few different translation problems that can be tackled with the tools we have developed, but the field is vast and much must be left to future work. The second part of this thesis is devoted to the study of one specific discourse phenomenon, the problem of pronominal anaphora. Pronominal anaphora is an intriguing object of study in that it is a fairly simple problem for a human language user, to the point that it might be considered uninteresting from the perspective of a human translator, yet it has an obvious potential to improve the quality of machine translation that has so far resisted all modelling attempts. Our contribution

related to this goal is the development of a cross-lingual pronoun prediction model to deal with pronominal anaphora in translation and its integration into our document-level SMT framework.

In the remainder of this introductory chapter, we address some loosely connected theoretical points concerning the relation between SMT and translation theory, the modelling assumptions underlying our experimental work and some considerations on the use of automatic evaluation methods. The purpose of these sections is to acquaint the reader with the foundations, assumptions and, more likely than not, prejudices that have influenced our research. This chapter also includes a section detailing the relation between this thesis and the corpus of previously published work on which it is based. In Chapter 2, we give an overview of the existing research literature on discourse in SMT to draw a picture of the relevant background.

The rest of the thesis is structured into two parts corresponding closely, but not exactly, to the two research goals outlined above. In the first part, we deal with the technical challenges of increasing the size of the context that feature models can take into account. We describe the solutions that have been proposed for document-level processing in SMT, introduce our new document-level decoding method and put it to the test with case studies on two discourse-level problems related to controlling the target language vocabulary used by the SMT system in different ways.

In the second part, we focus entirely on the translation of pronominal anaphora, a discourse-level problem that affects most SMT systems translating longer contiguous texts and cannot be solved correctly without some form of inference with access to document-level context. We discuss extensively what challenges the task of translating pronouns presents and describe an early approach to it. Then, we introduce a neural network classifier that models pronoun prediction as a separate task which is independent from the MT system. Finally, we conclude the second part by combining this classifier with the document decoding framework developed in the first part of the thesis and incorporating it as a feature function in the document-level decoder, uniting all the major contributions of our work in one single SMT system.

1.2 SMT and the Translation Process

The major part of this thesis and of the research it is based on follows the genre conventions of the SMT literature by adopting an engineering-oriented stance towards the problems we investigate. Beginning with the existing state of the art in SMT, which we have determined to be defective in certain aspects, we examine ways to capture some of these aspects with the proviso that all solutions must be realisable in the existing framework and can be subjected to immediate experimental scrutiny. Before we engage in this

pursuit, let us consider some fundamental contrasts between human translation activities and MT to shed some light on why it is difficult to deal with discourse-level text features in automatic translation.

The discourse-related limitations of SMT are to some extent technical and have to do with the necessity to constrain the search space of the MT system to ensure that the decoding problem remains computationally tractable. These aspects are discussed in some detail in Chapters 3 and 4 of this thesis. In addition to the technical constraints, however, there are conceptual limitations that make it difficult for an SMT system to acquire discourse competence.

In translation studies, the last century has brought about an important change of viewpoint, which has been named the *cultural turn* (Lefevere and Bassnett, 1995; Snell-Hornby, 2010). Until the last decades of the 20th century, translation was seen as an act of *transcoding* ("Umkodierung"), whereby elements of one linguistic sign vocabulary are substituted with signs of another linguistic sign vocabulary (Koller, 1972, 69–70). The principal constraint in this substitution is the concept of *equivalence* between the source language input and the target language output:

> Translating consists in reproducing in the receptor language the closest natural equivalent of the source-language message, first in terms of meaning and secondly in terms of style. (Nida and Taber, 1969, 12)

In the presentation of their theory of translation, Nida and Taber (1969, 12) emphasise that the primary aim of translation must be "reproducing the message", not the words of the source text. Their focus is on bible translation, so the word "message" in their writings strongly connotes the message of the gospel, but their theory is general enough to apply to other types of translation. According to them, translators "must strive for equivalence rather than identity" (Nida and Taber, 1969, 12). They stress the importance of *dynamic equivalence*, a concept of functional rather than formal equivalence that is "defined in terms of the degree to which the receptors of the message in the receptor language respond to it in substantially the same manner as the receptors in the source language" (Nida and Taber, 1969, 24). Koller (1972), primarily interested in general literary translation rather than bible translation, adopts a similar position. Instead of highlighting the message of the source text, he focuses on *understandability* and defines translation as the act of making the target text receptor understand the source text ("Übersetzen als Akt des Verstehbarmachens"; Koller, 1972, 67).

Equivalence as a purely linguistic concept has been criticised as deeply problematic because it fails to recognise the contextual parameters of the act of translating; it has even been called an "illusion" by Snell-Hornby (1995, 80), who also points out that the formal concept of equivalence "proved more suitable at the level of the individual word than at the level of the text" (Snell-Hornby, 1995, 80). The term is still used in a recent textbook on translation,

but, as the author points out, merely "for the sake of convenience" and "because most translators are used to it rather than because it has any theoretical status" (Baker, 2011, 5).

A key feature of more recent theoretical approaches to translation is their emphasis on the communicative aspects of translation. The cultural turn of the 1980s has been described to have "placed equivalence within a target-oriented framework concerned first and foremost with aspects of *target cultures* rather than with *linguistic* elements of *source* texts" (Leal, 2012, 43; her emphasis). Translation is seen as a "communicative process which takes place within a social context" (Hatim and Mason, 1990, 3). Instead of seeking for the target language text that is most closely equivalent to the source language input, the goal of translation is to perform an appropriate communicative act in the target community, and the target text is just a means of achieving this goal. Hatim and Mason (1990, 3) point out that doing so requires the study of *procedures* to find out "which techniques produce which effects" in the source and target community. According to them, texts are "the result of *motivated choice*" (Hatim and Mason, 1990, 4; their emphasis). In the case of translation, the motivations of the producer of the source text, as decoded by the translator, interact with the motivations of the translator him- or herself and determine the choices made to produce the target text.

Interestingly enough, when defending the novel way of understanding translation they promote, Lefevere and Bassnett (1995, 4) blame the shortcomings of previous theoretical approaches oriented towards linguistic equivalence on the influence of MT research and its demands for simple concepts that are easy to capture formally. Whether or not this explanation is true, it is striking how firmly even modern SMT techniques are rooted in traditional assumptions of translational equivalence and indeed how apt much of the criticism against such theories of translation is when applied to current standard methods in SMT.

The basis of all current SMT methods is the concept of word alignment, which was formalised by Brown et al. (1990, 1993) in the form still used today. Word alignments are objects of elaborate statistical and computational methods, but their linguistic meaning is defined simply by appealing to intuition:

> For simple sentences, it is reasonable to think of the French translation of an English sentence as being generated from the English sentence word by word. Thus, in the sentence pair (*Jean aime Marie*|*John loves Mary*) we feel that *John* produces *Jean*, *loves* produces *aime*, and *Mary* produces *Marie*. We say that a word is *aligned* with the word that it produces. Thus *John* is aligned with *Jean* in the pair that we just discussed. Of course, not all pairs of sentences are as simple as this example. In the pair (*Jean n'aime personne*|*John loves nobody*), we can again align *John* with *Jean* and *loves* with *aime*, but now, *nobody* aligns with both *n'* and *personne*. Sometimes, words in the English sentence of the pair align with nothing in the French sentence, and similarly, occasionally words in

> the French member of the pair do not appear to go with any of the words in the English sentence. (Brown et al., 1990, 80–81)

While this may indeed seem "reasonable" for simple sentences, the authors do not even try to elucidate the status or significance of word alignments in more complex sentences, where the correspondence between source and target words is less intuitive than in the examples cited. In practical applications, word alignments are essentially defined by what is found by the statistical alignment models used, and the issue of interpreting them is evaded completely. Even in articles dealing with manual word alignment and word alignment evaluation, it is not necessarily addressed (e. g., Lambert et al., 2005). While word alignments have been used in corpus studies aiming at a deeper understanding of the processes involved in translation (e. g., by Merkel, 1999), such efforts have had little impact on current practice in the SMT community.

The cross-linguistic relation defined by word alignments is a sort of translational equivalence relation. It maps linguistic elements of the source language to elements of the target language that are presumed to have the same meaning, or convey the same message. The same is true of the phrase pairs of phrase-based SMT (Koehn et al., 2003) and the synchronous context-free grammar rules of hierarchical SMT (Chiang, 2007), which are usually created from simple word alignments with mostly heuristic methods. None of these approaches exploits any procedural knowledge about linguistic techniques and their effects in the source and target community. Instead, it is assumed that each source text has an equivalent target text, possibly dependent on a set of context variables generally subsumed under the concept of *domain*, and that this target text can be constructed compositionally in a bottom-up fashion.

It is instructive to consider what type of translational equivalence can be accomplished with an SMT system. Clearly, nothing in current state-of-the-art SMT explicitly encourages dynamic equivalence. To model dynamic equivalence, an MT system would have to understand the purpose or function of the texts it translates, and there is no such knowledge in the existing models. However, one of the strengths of modern SMT is that it is capable of capturing correspondences that go beyond the simple word-by-word correspondences typical of pure formal equivalence. Often, SMT output can create quite a convincing illusion of dynamic equivalence, so we may consider that we are not doing justice to the SMT approach if we put it on the same level as simple literal translation. We know that real dynamic equivalence is beyond the scope of SMT models. An important factor is that the choice between competing translations suggested by the translation model in an SMT system is influenced to a large extent by the language model. The language model lacks all knowledge of the source text, which rules out the possibility of selecting target words as a function of the message or purpose of the input; it simply selects output words based on what has been observed most frequently in tar-

get language texts. Thus, we could say that an SMT system strives to achieve *observational equivalence* of the output with the input text.

In SMT, the notion of a domain is used to encode knowledge about the procedural aspects of translation referred to by Hatim and Mason (1990). Domain can be seen as a variable that all the probability distributions learnt by an SMT system are implicitly conditioned on, and it is assumed that if the domain of the system's training data matches the domain to which it will be applied, then the system will output contextually appropriate translations. If there is a mismatch between the training domain and the test domain, the performance of the system can be improved with domain adaptation techniques.

Although there is a great deal of literature on domain adaptation, few authors care to define exactly what a domain is. Frequently, a corpus of data from a single source, or a collection of corpora from similar sources, is referred to as a domain, so that researchers will refer to the "News" domain (referring to diverse collections of news documents from one or more sources such as news agencies or newspapers) or the "Europarl" domain (referring to the collection of documents from the proceedings of the European parliament published in the Europarl corpus; Koehn, 2005) without investigating the homogeneity of these data sources in more detail.

Koehn (2010, 53) briefly discusses the domain concept. He seems to use the word as a synonym of "text type", characterised by (at least) the dimensions of "modality" (spoken or written language) and "topic". Bungum and Gambäck (2011) present an interesting study of how the term is used in SMT research and how it relates to similar concepts in cognitive linguistics. In general, however, the term is used in a rather vague way and can encompass a variety of corpus-level features connected with genre conventions or the circumstances of text use. There is a clear tendency in current SMT to treat all aspects of a text either as very local, *n*-gram-style features that can easily be handled with the standard decoding algorithm or as corpus-level "domain" features that can conveniently be taken care of at training time.

According to Hatim and Mason (1990), human text production in general and translation in particular is a decision-making process involving a series of motivated choices. This is true also of SMT, where a decoding algorithm makes decisions based on some kind of formal utility measure parametrised by statistical models. Even the manner of text production can be quite similar. The most popular decoding algorithm for phrase-based SMT generates its output in natural reading order, pausing briefly every few words to deliberate on the next words to follow. This is precisely what a human translator might do when writing down a translation.

The difference between the human translator and the SMT system lies in the complexity of the decision-making process. Whenever it takes a decision, the SMT decoder has access to no more than a handful of words of context. Additionally, some general text-level word choice preferences may

be inscribed in the models in the form of "domain adaptation". By contrast, when pondering what words to choose to continue the same sentence, a competent human translator will have read and constructed a mental model of the whole text, will have talked to the commissioner of the translation about the target audience and the purpose of the translation, will have done additional research on the contents of the input text, will have made a text-level plan of the whole translation, will have mentally stored information used in making earlier decisions and will have thought about how to translate key passages in sentences to come. The context taken into consideration by the human translator exceeds that exploited by current SMT systems by far and includes knowledge about the whole document and its translation as well as background knowledge external to the document.

Given the current state of the art, we cannot hope to emulate the mental process of translation in its whole complexity, and we are far from formally modelling translation as a purposeful activity. With the work presented in this thesis, we strive to make a contribution towards removing the most basic restrictions on the size of the decision context in SMT and capturing some elementary discourse-level phenomena in translation with formal statistical models.

1.3 Modelling Assumptions

In developing the work described in this thesis, we have been guided by a set of assumptions that shaped the hypotheses we considered and explored. While there are good reasons to embrace these assumptions, we should point out that it is not a goal of this thesis to prove their validity, let alone their superiority over any other set of assumptions that could have been made. Rather, the principles outlined in the following paragraphs have a sort of axiomatic status in our work. They embody our endeavour to model linguistic phenomena in the way we consider most appropriate from a theoretical point of view rather than in the way that is most likely to result in quick gains, and an aversion to the principle of minimal incremental improvement, whose merit as a development strategy is undisputed, but which makes it difficult to explore any radical changes.

As a starting point, the models we develop are data-driven. This is a fairly uncontroversial assumption in the SMT community, even though it is not uncommon in production systems to include some components based on explicitly formalised linguistic knowledge. In our work, we avoid the creation of hard rules based on linguistic introspection. Instead, our goal is to use linguistic intuition along with corpus studies to create models whose parameters can then be estimated from data. We believe that this type of model is more versatile and has greater flexibility to deal with corpus data that may not always match the educated human's idea of grammaticality.

Taking our reliance on raw corpus data even further, we aim to develop models that depend as little as possible on explicit annotations. Corpus annotation is another way to encode introspective linguistic knowledge. In many subfields of natural language processing (NLP), it is common to enrich corpus data with explicit annotations reflecting a phenomenon of interest and then train statistical models on this data. This approach is usually considered to be fully data-driven, since it relies on data sets sampled from real corpora, reflecting the distribution of texts attested in everyday linguistic production. Nevertheless, explicit annotation always imposes a certain underlying structure on a text, and it is difficult to ensure that the selected structure optimally reflects the information needed in a translation scenario. This is why we have a preference for models that manipulate raw text data, even though we do depart from this principle and use a part-of-speech tagger or an anaphora resolution system trained on annotated data in some cases.

Rather than working with explicitly annotated data or proceeding in a completely unsupervised way, we attempt to use the information contained in parallel bitexts instead. This is the one type of high-quality annotations that is abundant in an SMT setting. Much of the parallel text included in typical SMT training corpora is created by expert translators with high quality standards. It contains a wealth of information and is available in very large quantities compared to other types of annotations, but the translators creating the bitexts were ignorant of how their texts would later be used in a computational setting. As a result, the annotations we have are completely unbiased towards our own purposes. This makes the annotations potentially noisy and difficult to use, but it also ensures that they are representative samples of distributions encountered in real-life translations, which should contribute to the validity of the models we derive from them.

Finally, it has been a goal in our work to give preference to integrated approaches over pipeline solutions and to enable joint inference over multiple steps wherever possible. While pipeline approaches make it easy to decompose a task into small manageable steps, they have a tendency towards developing complex dependencies between the individual steps and propagating errors from one step to the next. This is why we implement document translation as a part of the core SMT decoding process (Chapter 4) rather than performing inference on word lattices or n-best lists output by a standard decoder, and it is why we model anaphora resolution and pronoun prediction jointly in a single neural network (Chapter 8).

1.4 MT Evaluation and Translation Quality

Nobody performing experiments on MT can evade the question of evaluation. For practical reasons, MT quality is usually measured with automatic metrics such as BLEU (Papineni et al., 2002), which match word sequences in

the translated text against reference translations produced by human translators and assume that greater overlap is correlated with higher translation quality. The inadequacy of metrics of this type is widely recognised and acknowledged, but few reasonable alternatives are available, and none of them is generally accepted.

A key problem for the development of high-quality MT is the fact that the very concept of translation quality is not well-defined. Human evaluation of translations, the gold standard for all translation quality measurement, is a highly non-trivial task in itself. A human translator who renders a text in another language makes a great number of choices to select appropriate words in the target language. To some extent, these choices are guided by the wording of the input text, but they also depend on various extra-linguistic factors such as the proposed use and target audience of the translation, cultural background knowledge of the communities for which the source and target texts are written, language-specific genre conventions, economic considerations, media-specific constraints, etc. A reasonable method to evaluate a translation is to make assumptions about such context factors and to discuss the adequacy of the decisions taken by the translator in the light of the assumptions made.

This intellectual approach to translation criticism may work well for the education of human translators, but it is defeated in MT research not only by its extreme cost, but also by several other factors impairing its usefulness. Essential evaluation parameters such as target audience and intended use are often ill-defined in MT research. The markedly non-intellectual translation process embodied in an SMT system and the sheer difficulty of exploiting the insights gained by such a process render translation criticism unsuitable as a tool for MT development. As a result, it is usually substituted by sampling methods where humans are asked, e. g., to rank a number of translations (often single sentences with very little context) by quality. By measuring inter-annotator and intra-annotator agreement, the reliability of such methods can be assessed to some extent, but it is next to impossible to prove their validity since the precise evaluation criteria are often left to the evaluators' intuition (explicitly so, e. g., by Callison-Burch et al., 2012, 14). However, as Artstein and Poesio (2008, 557) point out, agreement between evaluators does not entail validity because "[t]wo observers of the same event may well share the same prejudice while still being objectively wrong." Moreover, even if the evaluators have objectively sound reasons to prefer one disfluent translation over another, their judgements are influenced by effects of salience and some errors go unpunished more easily than others, although they do reflect fundamental problems of the generating MT system.

The development of automatic MT evaluation metrics is an object of ongoing research. For more than a decade, BLEU (Papineni et al., 2002) has been the standard metric in MT research. BLEU considers the overlap of n-gram sequences between the candidate translation and one or more reference trans-

lations. It consists of two components. The first represents *n*-gram precision in the candidate translation, which is defined as the number of *n*-grams the candidate shares with the reference divided by the total number of *n*-grams in the candidate translation. This quantity is computed for 1-grams to 4-grams and aggregated into a geometric mean. It is then multiplied with the second component, a brevity penalty which assumes the function of a recall measure. The brevity penalty punishes translations with a factor that decays exponentially with the length ratio between candidate and reference translation if the candidate translation is shorter than the reference.

BLEU has been used both for assessing the quality of MT systems and as an objective function for automatic parameter tuning (Och, 2003). Significant research efforts have been spent on improving BLEU scores. By its nature, BLEU favours locally fluent MT output, and advances in *n*-gram language modelling methods often have large impact on BLEU. By contrast, long-range dependencies are not captured, and discourse-level phenomena are reflected much less reliably by the metric.

Since the introduction of BLEU, many other metrics have been proposed. None of them has been able to replace BLEU as the standard metric, but some of them have gained some popularity. Among the more popular alternatives, we could mention NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005; Denkowski and Lavie, 2011) and TER (Snover et al., 2006). While these metrics address some of the shortcomings of BLEU, they do not add any specific support for discourse-level phenomena. Some discourse-level MT evaluation measures have recently been suggested (Giménez et al., 2010; Wong et al., 2011; Wong and Kit, 2012; Guzmán et al., 2014; Joty et al., 2014), but they have been developed and tested for English as a target language only, whereas English is the source language in most of the experiments discussed in this thesis.

MT evaluation is an interesting research problem in itself, but it is not a focus of our work. However, in experimental work it cannot be avoided completely. Our stance on evaluation is to adopt standard evaluation measures and, in particular, the BLEU score, while recognising their inadequacy. We generally report BLEU scores for all experiments, but we do not necessarily expect that they reliably reflect the quality of discourse-level features in the translation. In Chapter 7, we introduce an automatic evaluation metric that gauges the accuracy of pronoun translation more specifically than standard evaluation measures do, but it suffers from many of the same shortcomings as the existing methods. Currently, the only method that has some claim to validity when it comes to measuring discourse-level features of translation is a targeted human evaluation like the one we conduct for our SMT experiments in Chapter 9.

As a result of these considerations, we do not generally perform statistical hypothesis tests involving BLEU scores or similar metrics. Hypothesis tests serve to prove that a difference between two observed measurements is

unlikely to be due to chance, suggesting that it reflects a substantive change in the experimental outcome. However, since we have serious doubts about whether the measurements we consider actually reveal the qualities we are most interested in, this is immaterial for score differences small enough that their significance can be called in doubt. In any case, we cannot draw reliable conclusions from them, and labelling them as significant would confer a false sense of importance to them. We therefore do indicate BLEU scores following standard practice in the research community, but as we consider the validity of the scores to be a more serious concern than their significance, we do not attempt to prove significance formally.

1.5 Relation to Published Work

Much of the material contained in this thesis has been published previously, primarily in the form of conference papers. The text of the published papers was used, in updated and extended form, as the basis for various parts of the thesis. In particular, the individual chapters are related to prior publications as follows:

- An earlier version of the literature survey in *Chapter 2* was published as an article in the journal *Discours* (Hardmeier, 2012).
- The decoding procedure discussed in *Section 3.4* was described in a paper presented at the International Workshop on Spoken Language Translation (IWSLT) in Paris, France, 2010 (Hardmeier and Federico, 2010).
- The document-level decoding algorithm proposed in *Chapter 4* was published in a paper presented at the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) on Jeju Island, Korea, 2012 (Hardmeier et al., 2012).
- Our software implementation of this algorithm, the Docent decoder, was presented at the system demonstration session of the 51st Annual Meeting of the Association for Computational Linguistics (ACL) in Sofia, Bulgaria, 2013 (Hardmeier et al., 2013a). I implemented all the core functionality of the Docent decoder and wrote the larger part of the system description paper, with the exception of a section on readability models written by Sara Stymne.
- The results on document-level feature weight optimisation in *Section 4.7* were published in a paper presented at the Workshop on Discourse in Machine Translation (DiscoMT) in Sofia, Bulgaria, 2013 (Stymne et al., 2013a). The experimental work leading to these results was carried out by Sara Stymne, who also composed the text of the DiscoMT paper. I

participated in the discussions leading to the experiments and contributed some advice on practical issues related to the Docent decoder.

- A description of the semantic document language model described in *Section 5.1.3* was included as a part of our paper at EMNLP-CoNLL 2012 (Hardmeier et al., 2012).
- The work on readability in *Section 5.2* was published in a paper presented at the 19th Nordic Conference of Computational Linguistics (NODALIDA) in Oslo, Norway, 2013 (Stymne et al., 2013c). The experiments in this work were carried out by Sara Stymne, who also composed the text of the NODALIDA paper. I participated in the discussions leading to the experiments and contributed some advice on practical issues related to the Docent decoder.
- The corpus study on pronoun translation in *Section 6.3* was a part of our paper at IWSLT 2010 (Hardmeier and Federico, 2010).
- An earlier version of the discussion of challenges in pronoun translation in *Section 6.4* was included as a part of the *Discours* article referred to above (Hardmeier, 2012).
- The word dependency model and the pronoun evaluation metric described in *Chapter 7* were included in our paper at IWSLT 2010 (Hardmeier and Federico, 2010).
- The cross-lingual pronoun prediction model in *Chapter 8* was published as a paper at the Conference on Empirical Methods in Natural Language Processing (EMNLP) in Seattle, USA, 2013 (Hardmeier et al., 2013b).
- The SMT system with anaphora handling described in *Chapter 9* was used in our submission to the shared task on English–French MT at the Ninth Workshop on Statistical Machine Translation in Baltimore, USA, 2014, and was discussed in a system description paper (Hardmeier et al., 2014). In the system description paper, I was responsible for the experimental work as well as the description of the English–French SMT system.

The vast majority of the results in this thesis are joint work with my advisors Joakim Nivre and Jörg Tiedemann (Uppsala University) and Marcello Federico (Fondazione Bruno Kessler, Trento). Their contributions are not marked separately. Except as mentioned otherwise above, the main responsibility for the complete scientific process from conception and experimentation to analysis and writing was mine.

2. Research on Discourse and SMT

The importance of discourse-level dependencies for translation has only recently attracted systematic attention in the SMT community. In a survey paper about discourse in SMT published only a short time ago, we pointed out the "SMT community's apparent lack of interest in discourse" (Hardmeier, 2012) and showed that most of the research on discourse-related problems in SMT was conducted under different headings such as terminological consistency or domain adaptation. Since then, the number of papers explicitly interested in discourse has grown, and there was even an ACL workshop devoted to this topic (DiscoMT 2013 in Sofia, Bulgaria). There are different strands of research in the literature. One attempts to exploit the macroscopic structure of the input texts to infer better translations. Some work is concerned with different aspects of lexical cohesion, terminological consistency and word choice. Other work deals with specific linguistic features that are governed by discourse-level processes such as generation of anaphoric pronouns, translation of discourse connectives or verb tense selection. Yet another strand addresses the technical challenges involved in processing document-level information and seeks to create a software infrastructure that straightforwardly supports discourse-level translation. In this chapter, we review and discuss the existing literature.

2.1 Discourse Structure and Document Structure

One of the earliest attempts to integrate discourse processing into SMT is also, in a sense, one of the most ambitious. Several years before the phrase-based (Koehn et al., 2003) and hierarchical (Chiang, 2007) approaches to SMT were introduced, Marcu et al. (2000) suggested doing MT by rewriting discourse structure trees. They compared the discourse structure of a small corpus of Japanese and English parallel documents and concluded that "if one attempts to translate Japanese into English on a sentence-by-sentence basis, it is likely that the resulting text will be unnatural from a discourse perspective" (Marcu et al., 2000, 12–13) because of significant structural differences at the sentence, paragraph and text levels. They outline a discourse transfer model to rewrite the discourse structure of an input text into a corresponding tree for the target language. To our knowledge, this work has never been followed up after its initial publication, and we are not aware of any actively developed SMT system implemented along these lines.

In the more recent SMT literature, there is some work on exploiting text-level structure for specific text genres. Foster et al. (2010) perform local language model (LM) adaptations in a system translating Canadian parliamentary debates using metadata features that represent various aspects of document structure. Wäschle and Riezler (2012) apply a multi-task variant of minimum error-rate training (Och, 2003) to fine-tune their models to different text sections in patent translation. Louis and Webber (2014) improve the translation of biographical texts in Wikipedia with a cache LM influenced by a topic model that can account for the blockwise topic shifts typical of this text genre.

2.2 Cohesion, Coherence and Consistency

Lexical choice is a problem that has traditionally attracted much attention in SMT research. Initially most studied from the points of view of language modelling and domain adaptation, the effects of text-level features on word choice have recently moved into focus. The linguistic key concepts are cohesion and coherence, two fundamental discourse properties that establish "connectedness" in a text (Sanders and Pander Maat, 2006, 591). *Cohesion* is a surface property of the text that is realised by explicit clues such as the use of discourse markers or word repetition. It occurs whenever "the interpretation of some element in the discourse is dependent on that of another" (Halliday and Hasan, 1976, 4). *Coherence*, by contrast, is related to the connectedness of the "mental representation of the text rather than of the text itself". It is created referentially, when different parts of a text refer to the same entities, and relationally, by means of coherence relations such as *Cause–Consequence* between different discourse segments (Sanders and Pander Maat, 2006, 592).

Another term that has sometimes been used by less linguistically oriented researchers is that of *lexical* or *terminological consistency*. The underlying assumption is that the same concepts should be consistently referred to with the same words in a translation. To what extent this principle holds in naturally occurring texts of different genres, and to what extent and in what ways it is or should be enforced in SMT systems, is an object of ongoing research.

2.2.1 Corpus Studies

In computational word sense disambiguation software, it is common, and usually beneficial, to impose a *one sense per discourse* constraint (Gale et al., 1992) and assume that all uses of a polysemous term in the same document denote the same sense of that term. Carpuat (2009) investigates a similar *one translation per discourse* hypothesis that relates to translated texts, supposing that all instances of the same term in a document should be translated in the same way. By examining human reference translations for two English–French

SMT test sets, she finds indeed that 80 % of the French words are aligned to no more than one English translation and 98 % to at most two translations, after lemmatising both source and target. Looking at machine translations of the same test sets, she observes that the regularity in word choice is even stricter in SMT as a result of the generally low lexical variability of SMT output.

These results suggest that there is not much to be gained by just enforcing consistent vocabulary choice in SMT, since the vocabulary is already fairly consistent. In principle, it may be possible to improve SMT by using whole-document context to select translations. However, a more recent study by Carpuat and Simard (2012) shows that this may be more difficult than it seems. In that study, the authors find consistency and translation quality to be essentially uncorrelated or even negatively correlated in SMT output. In particular, they show that machine-translated output tends to be more consistent when produced by systems trained on smaller corpora, indicating that "consistency can signal a lack of coverage for new contexts" rather than being a sign of translation quality (Carpuat and Simard, 2012, 446). In a manual analysis of post-edited MT output, they find that most lexical inconsistencies are symptoms of more fundamental problems such as outright semantic translation errors or syntactic or stylistic problems, whereas the terminological inconsistencies typically found in imperfect human translations only account for about 13–16 % of the inconsistent translations. These findings are encouraging in the sense that, in the best case, a model improving MT output consistency in the right way might help to fix some of the more fundamental errors as well, but the lack of positive correlation between measured consistency and translation quality shows that it is important to enforce not only consistent, but also correct translations, and that it may be necessary to make use of additional information for good results.

The *one translation per discourse* hypothesis is tested again by Ture et al. (2012), using a methodology based on forced decoding with a hierarchical SMT system and examining the translations selected by human translators at text positions where multiple options would have been available in the SMT rule table. They find that the human translators indeed opt for consistent lexical choice in the majority of cases, but that some content words may be translated in more varied ways because of stylistic considerations. They propose a set of cross-sentence feature functions rewarding translation rule reuse that achieves significant improvements in Arabic–English and Chinese–English translation tasks.

Another corpus study about lexical cohesion in MT output was published by Voigt and Jurafsky (2012). They compare referential chains in a literary text and a piece of news text in Chinese with their English translations generated by the on-line MT service Google Translate. In the source language, both texts exhibit a similar number of entities, but the referential chains in the literary text are denser, indicating stronger cohesion, and contain more pro-

nouns. They find the MT system to be relatively successful at transferring these chains to the target language. For the news text, the characteristics of the referential chains in the output are similar to the statistics of human translations; for the literary text, there is a slight tendency towards underexpression of cohesive devices.

In a study investigating lexical consistency in human translations and machine translations of texts in different genres, Guillou (2013) observes that the lexical consistency of human translations varies across word classes. For most of her texts, the consistency of noun translations is fairly high, but not perfect. For verbs, there is greater variation. In particular, the most common verbs belonging to the top 5 % when ordered by frequency are translated much less consistently. Guillou therefore concludes that consistency is not invariably desirable and should be enforced only selectively. In machine-translated texts, she finds, in accordance with Carpuat and Simard (2012), that the measured consistency is high on average, but this does not necessarily mean that the translations are correct. Disambiguation of polysemous words is a serious problem for an SMT system, and document-level consistency is often insufficient as a predictor of translation quality. An important difference between human translations and machine translations is that inconsistencies in the former often just represent different wordings of the same notions, whereas incorrect word choices made by SMT systems can completely distort the meaning of the translation and have a serious impact on the adequacy of the translations.

Beigman Klebanov and Flor (2013) examine the vocabulary distribution of translated texts in terms of "associative texture". The objective measure used by their study is the "word association profile", defined as the distribution of pointwise mutual information between pairs of content word types in a text, and the mean of this distribution, called "lexical tightness". The authors find that lexical tightness is systematically and significantly lower in texts that were machine-translated into another language and back again than in the original input texts. It is also lower in MT output than in human reference translations, and it is lower in machine translations of lower quality than in better machine translations, where translation quality is determined by human evaluation.

2.2.2 Cross-Sentence Language Models

One way to promote cohesive lexical choice across sentence boundaries is to extend the scope of the language model history by propagating information between sentences. Tiedemann (2010a,b) suggests using an exponentially decaying cache to carry over word preferences from one sentence to the next. He demonstrates modest improvements with this approach with a corpus of medical texts (Tiedemann, 2010a), while the same technique fails when ap-

plied to newswire text (Tiedemann, 2010b). One significant problem is that the cache easily gets contaminated with noise, and that it can contribute to the propagation of bad translations to the following sentences. More recently, improvements have been demonstrated with a more sophisticated caching technique that initialises the cache with statistics from similar documents found with information retrieval methods and keeps the noise level in check with the help of a topic model created with Latent Dirichlet Allocation (LDA; Gong et al., 2011a). A cache model presented by Louis and Webber (2014) is similar, but extends the topic model with the capacity to detect topic shifts to account for the semi-structured nature of the texts translated (biographic articles from Wikipedia).

As the requirements on translational consistency vary across word classes (Guillou, 2013), it can make sense to create a model covering only the words that are most susceptible to benefit from cohesion modelling. This is what we have attempted to do with a cross-sentence semantic space *n*-gram model over content words (Hardmeier et al., 2012). This model is described in more detail in Section 5.1.3.

2.2.3 Lexical Cohesion by Topic Modelling

Some researchers have proposed methods based on Latent Semantic Analysis (LSA) and LDA to achieve lexical cohesion under a topic model. Kim and Khudanpur (2004) use cross-lingual LSA to perform domain adaptation of language models in one language (assumed to suffer from sparse resources) given adaptation data in another language. Zhao and Xing (2006) present an approach to word alignment named BiTAM based on bilingual topic models, which they then extend to cover SMT decoding as well (Zhao and Xing, 2008). A similar technique based on a bilingual variant of LDA is used by Tam et al. (2007) for adapting language models and phrase tables.

Simpler and more recent approaches include the one by Gong et al. (2010), who adapt SMT phrase tables with monolingual LDA, and Ruiz and Federico (2011), who implicitly train bilingual LSA topic models by concatenating short pieces of text in both languages before training the model, and use these topic models for language model adaptation. Gong et al. (2011b) use *n*-best rescoring to make the topic distribution for each document as similar as possible to the corresponding distribution in the source document, achieving a marginal improvement in a Chinese–English task. Eidelman et al. (2012) adapt features in the phrase table based on an LDA topic model. They compare adaptation at the sentence level with per-document adaptation and find that, while both approaches work, sentence-level adaptation gives marginally better results on their Chinese–English tasks. Hasler et al. (2014) completely integrate LDA with phrase table training by estimating phrase translation

probabilities with a bilingual LDA model which directly represents parallel documents as bags of phrase pairs.

2.2.4 Encouraging Lexical Consistency

There have been several attempts directly aimed at improving the consistency of lexical choice in the MT output. Xiao et al. (2011) present a two-pass decoding approach to enforce consistent translation of recurring terms in a document in Chinese–English newswire translation. After the first pass, they disambiguate terms with multiple translations by finding the dominant translation in an n-best list. Then they filter the phrase table of the second decoding pass to remove inconsistent translations. Their research is followed up by the work by Ture et al. (2012) cited above, which realises improvements for Chinese–English and Arabic–English by designing features to guide the second-pass translation process instead of manipulating the phrase table as Xiao et al. (2011) do.

Alexandrescu and Kirchhoff (2009) describe a graph-based learning approach to favour similar translations for similar input sentences by considering similarity both between training and test sentences and between pairs of test sentences, which leads to large improvements for Italian–English and Arabic–English SMT tasks.

Ma et al. (2011) argue that the consistency of translations can be improved by constraining SMT output to be similar to sentences retrieved from a translation memory. However, their method does not explicitly enforce or encourage cross-sentence consistency. Instead, they entirely rely on the assumption that the examples supplied by the translation memory will be more consistent than what the SMT system would generate on its own.

2.2.5 Models of Cohesion and Coherence

Xiong et al. (2013b) describe a model that explicitly tries to capture the notion of lexical cohesion in Chinese–English SMT. Their basic model scans the output of their MT system for *lexical cohesion devices*, which are pairs of target language words satisfying certain cohesive relations. The relations considered are identity (word repetition), synonymy or approximate synonymy and hyponymy or hypernymy; they are detected with the help of WordNet (Fellbaum, 1998). The authors show that significant gains in MT quality can be realised just by rewarding the occurrence of such cohesion devices. Scoring them with more sensitive metrics based on conditional probability and mutual information increases the gain. Similar effects can be achieved by considering *bilingual cohesion triggers* formed by replacing the first one of the words in a lexical cohesion device with the source language words aligned to it (Ben et al., 2013).

Instead of considering isolated word pairs, lexical cohesion can be modelled by looking at chains of words extending through the whole document. Xiong et al. (2013a) start by identifying lexical chains in the source language with a thesaurus-based algorithm (Galley and McKeown, 2003). Next, they map the lexical chains into the target language with a set of maximum entropy classifiers predicting the best translation of a source word given both its local context and the neigbouring words in the chain. Finally, they add a feature model to their hierarchical SMT decoder to encourage it to adopt the word choices predicted by the classifiers. This model improves translation quality substantially over the word pair models. In a variant of this model, Xiong and Zhang (2013) use a Hidden Topic Markov Model (Gruber et al., 2007) instead of the thesaurus-based lexical chain extractor to generate chains of semantically related words.

2.3 Targeting Specific Discourse Phenomena

In contrast to the models described in the previous section, which are concerned with lexical cohesion and word choice in a quite general sense, there have been recent efforts to develop models dealing with the realisation of distinct types of cohesive relations. Often, such relations are specifically encoded with particular word classes. The problems that have been studied include the correct translation of anaphoric pronouns, the generation of determiners in noun phrases, tense marking on verbs and the translation of discourse connectives.

2.3.1 Pronominal Anaphora

Pronominal anaphora is the use of a pronoun to refer to an entity mentioned earlier in the discourse. This happens very frequently in most types of connected text. This phenomenon will be the main topic of the second part of this thesis, where our own results are discussed in great detail.

Usage and distribution of pronouns differ between languages (Russo et al., 2011). When an anaphoric pronoun is translated into a language with gender and number agreement, the correct form must be chosen according to the gender and number of the translation of its antecedent. Corpus studies have shown that this can be a problem for both statistical and rule-based MT systems, resulting in a potentially large number of mistranslated pronouns depending on language pair and text type (Hardmeier and Federico, 2010; Scherrer et al., 2011).

It was recognised years ago that the information contained in parallel corpora may provide valuable information for the improvement of anaphora resolution systems, but there have not been many attempts to cash in on this insight. Harabagiu and Maiorano (2000) exploit parallel data in English and

Romanian to improve pronominal anaphora resolution by merging the output of anaphora resolvers for the individual languages with a set of simple rules. Mitkov and Barbu (2003) pursue a similar approach for English and French. They create a more elaborate set of handwritten rules to resolve conflicts between the output of the language-specific resolvers. Veselovská et al. (2012) resolve different uses of the pronoun *it* in English–Czech data with handwritten rules that benefit from both monolingual and bilingual features. Other work has used word alignments to project coreference annotations from one language to another with a view to training anaphora resolvers in the target language (Postolache et al., 2006; de Souza and Orăsan, 2011). Rahman and Ng (2012) instead use MT to translate their test data into a language for which they have an anaphora resolver and then project the annotations back to the original language.

The converse problem, exploiting anaphora information for the improvement of SMT systems, was first addressed by Le Nagard and Koehn (2010). They approach the translation of anaphoric pronouns in phrase-based SMT by processing documents in two passes: The English input text is run through a coreference resolver developed by the authors *ad hoc*, and translation is performed with a regular SMT system to obtain French translations of the antecedent noun phrases. Then the anaphoric pronouns of the English text are annotated with the gender and number of the French translation of their antecedent and translated again with another MT system whose phrase tables are annotated in the same way. This does not result in any noticeable increase in translation quality, a fact which the authors put down to the insufficient quality of their coreference resolution system. However, in a later application of the same approach to an English–Czech system, no clearly positive results are obtained despite the use of data manually annotated for coreference (Guillou, 2011, 2012).

Engaging in the same task, Hardmeier and Federico (2010) create a one-pass system that directly incorporates the processing of coreference links into the decoding step. This system is described in Chapter 7. Pronoun coreference links are annotated with the BART anaphora resolution software (Versley et al., 2008). We then add an extra feature to the decoder to model the probability of a pronoun given its antecedent. Sentence-internal coreference links are handled completely within the SMT dynamic programming algorithm. For links across sentence boundaries, the translation of the antecedent is extracted from the MT output after translating the sentence containing it, and it is held fixed when the referring pronoun is translated. In that work, no improvement in BLEU score is achieved for English–German translation, but a slight improvement is found with an evaluation metric targeted specifically to pronoun coreference. A subsequent attempt to apply the same technique to the language pair English–French is largely unsuccessful (Hardmeier et al., 2011). In later work, we model anaphoric relations discrim-

inatively with neural network classifiers (Hardmeier et al., 2013b). This work and its application to SMT is described and discussed in Chapters 8 and 9.

For the Czech language, there is a body of research in the TectoMT framework (Žabokrtský et al., 2008), which combines deep syntactic analysis with statistical transfer methods. Novák (2011) investigates the performance of the TectoMT system on translating the English pronoun *it* into Czech. He presents an analysis of errors made by the MT system and finds that about half of the occurrences of the pronoun *it* in his corpus are non-referring expletives or refer anaphorically to constituents that are not noun phrases. In such cases, the obvious translation of *it* with a Czech neuter pronoun is most often correct. The pronoun is also consistently translated with a Czech neuter when it does have noun phrase (NP) reference, and a substantial part of these cases are wrong. Novák et al. (2013a) suggest using a discriminative classifier with features derived from the tectogrammatical structure to predict the morphological features of translations of *it*. Even though their classifier beats an uninformed baseline by a large margin, there is no effect on BLEU. Manual evaluation shows that the changes with respect to the baseline correspond to improvements somewhat more often than to degradations. In later work, this approach is extended to reflexive pronouns (Novák et al., 2013b). For reflexives, the improvements in the manual evaluation are more consistent, but the BLEU scores are still unaffected.

Russo et al. (2012a,b) address a somewhat different problem. They consider the generation of subject pronouns when translating from pro-drop languages into languages that require pronominal subjects to be realised explicitly, conducting a corpus study and examining the output of a rule-based and a statistical MT system. Their work focuses on identifying where to insert pronouns with the help of rule-based preprocessing and a statistical postprocessing step. They do not make any attempt to resolve pronominal anaphora and resort to inserting majority class (masculine) pronouns whenever there is an ambiguity. By doing so, they manage to improve the pronoun translation accuracy of their rule-based translation system.

Taira et al. (2012) test the impact of inserting explicit pronouns for implicit subjects and objects in Japanese on phrase-based SMT into English. They manually insert pronouns into Japanese source sentences in contexts where they are not required by Japanese grammar, but would be required in a corresponding English sentence. After SMT into English, they observe only a marginal improvement in BLEU score, but a larger gain with an *ad hoc* metric sensitive to this specific phenomenon.

Since automatic anaphora resolution is difficult and error-prone, it is of great value for the development of anaphora-aware SMT systems to have test corpora manually annotated for coreference. The standard corpora used in anaphora resolution research are often insufficient because they are only available in one language. Harabagiu and Maiorano (2000) mention translating some of the coreference-annotated training data of the Message Under-

standing Conferences (MUC; Grishman and Sundheim, 1996) into Romanian, but we do not know if this translation is publicly available. Recently, coreference annotations have been added to a number of parallel corpora. These include the Prague Czech–English Dependency Treebank (PCEDT; Hajič et al., 2006) with parallel text in English and Czech and the ParCor pronoun coreference corpus (Guillou et al., 2014) with parallel text in English and French as well as English and German. The Copenhagen Dependency Treebank (Buch-Kromann et al., 2009) supposedly contains annotated parallel text for Danish and English, Italian, Spanish and German, but it is unclear if and when these annotations will be completed and released. An English–French data set released by Popescu-Belis et al. (2012b) contains a reduced form of pronoun annotations, labelling the English pronouns with the word they correspond to in the French text, but not actually marking their antecedents.

2.3.2 Noun Phrase Definiteness

Definiteness marking of noun phrases is a phenomenon governed by non-trivial language-specific discourse features that vary even among closely related languages. In some languages like Russian or Czech, noun phrases have no overt morphological definiteness markers. When translating from these languages into a target language like English that requires the use of definite or indefinite articles, a standard SMT system will not have the necessary information to generate correctly distributed definite and indefinite articles.

Knight and Chander (1994) describe a statistical postediting system based on decision tree classifiers for inserting English definite and indefinite articles into the output of rule-based MT systems. They also make an experiment with human informants to determine how much discourse information is necessary to solve this task. When presented with isolated noun phrases extracted from a corpus, their subjects decide correctly if the noun phrase was definite or indefinite in the corpus in around 80 % of the cases, clearly exceeding the simple majority class accuracy of 67 %. When given access to discourse context, the human annotators' accuracy reaches 95 %. These figures give an indication of the upper bounds on accuracy that can be achieved in such a task.

Tsvetkov et al. (2013) extend SMT systems for Russian–English and Czech–English with a classifier to predict NP definiteness trained on sentence-level lexical and morphosyntactic features. To make sure that the required phrases are available to the MT system, they enrich their phrase tables with synthetic phrase pairs containing unseen determiner-noun pairs. They demonstrate that this technique improves BLEU scores with respect to a standard baseline and that it compares favourably to a determiner insertion procedure at post-processing time.

2.3.3 Verb Tense and Aspect

Gong et al. (2012b) present a cross-sentence model to control the generation of correct verb tenses in the MT output. This is a problem that occurs in the translation from Chinese to English because Chinese verbs are not morphologically marked for tense, whereas generating correct English output requires selecting the right tense form. They use *n*-gram-like features on the target side to model the English sequence of tenses, with two different models to capture the sequence of verb tenses within a sentence and across sentences, respectively. Their cross-sentence model is just a sequence model over the tenses of the main verbs in each sentence. Sentences are processed in order, and information about the tense of the main verb generated is passed on to the following sentences so that the tense of the next verb can be conditioned on this information. By applying this model, they achieve sizeable improvements in BLEU on a Chinese–English task.

One weakness of the *n*-gram tense model is that it only incorporates target language information. Gong et al. (2012a) achieve additional improvements by replacing the *n*-gram model with a support vector machine classifier exploiting both source language and target language features. Furthermore, they expand the phrase table with synthetic entries to ensure that all required verb forms are available to the SMT system.

Meyer et al. (2013) explore a related problem in English–French translation. Owing to differences in the aspect marking systems of English and French, an English simple past verb can correspond to an *imparfait*, *passé simple* or *passé composé* form in French. A key property for predicting this distinction is called *narrativity*. Meyer et al. (2013) train a classifier to predict the narrativity of English past tense verbs. They show that a small improvement in BLEU scores and a beneficial effect in manual evaluation can be achieved by integrating the narrativity feature in a factored phrase-based SMT system (Koehn and Hoang, 2007).

2.3.4 Discourse Connectives

The translation of discourse connectives has recently been studied as a main focus of the Swiss COMTIS project on text-level SMT (Popescu-Belis et al., 2012a), which resulted in a number of publications on this topic. In a corpus study, Cartoni et al. (2011) compare parts of the Europarl multilingual corpus (Koehn, 2005) that were originally written in French with other parts translated into French from English, German, Italian and Spanish. They find that the different subcorpora use fairly similar vocabulary in general, but that discourse connectives have significantly different distributions depending on the original source language of the text. They also notice that it is fairly common for translators to introduce discourse connectives not explicitly found in the source language, and less common to leave out connectives present

in the source. Meyer et al. (2011b) contrast findings from a corpus study based on manual annotation with results obtained from the exploration of parallel corpora. Detailed results of the study are not contained in the published abstract. Meyer and Webber (2013) study the translations of discourse connectives from English into French and German and find that up to 18 % of explicit English discourse connectives have no direct correspondence in French or German human translations, whereas machine translations much more often include literal translations of connectives.

Without any relation to the COMTIS project, Becher (2011a,b) studies implicitation and explicitation of discourse connectives in a descriptive corpus study of business texts translated between German and English. He approaches these phenomena from the angle of translation studies rather than natural language engineering and proposes explanations in terms of features of the grammatical systems of the source and target language and in terms of properties of the translation process.

Meyer et al. (2011a) and Meyer (2011) investigate automatic disambiguation of polysemous discourse connectives. They propose a "translation spotting" annotation scheme for corpus data that marks up words that can be translated in different ways with their correct translation, which they call "transpot", instead of explicitly annotating linguistic features (Popescu-Belis et al., 2012b; Cartoni et al., 2013). Disambiguating connectives with an automatic classifier before running a phrase-based SMT systems results in small improvements in translation quality for English–French (Meyer, 2011; Meyer and Popescu-Belis, 2012; Meyer et al., 2012) and English–Czech (Meyer and Poláková, 2013) according to some *ad hoc* evaluation criteria, even though the BLEU scores are largely unaffected. Meyer et al. (2012) present a family of automatic and semi-automatic evaluation scores called ACT to measure the accuracy of discourse connective translation in order to obtain a more meaningful assessment of progress on this problem than what a general-purpose measure like BLEU can deliver. These metrics are then further studied and validated against human judgements for the language pairs English–French and English–Arabic (Hajlaoui and Popescu-Belis, 2012, 2013).

2.4 Document-Level Decoding

In standard SMT systems, it is relatively difficult to exploit discourse-level features because of the limitations of the decoding algorithm. Phrase-based SMT decoders almost universally use a variant of the dynamic programming beam search algorithm described by Koehn et al. (2003) for decoding. This algorithm combines good search performance with high efficiency thanks to a dynamic programming technique exploiting the locality of the models, making it difficult or impossible to integrate models whose dependencies require considering a context larger than a window of five or six words. In

past research, this problem was addressed mostly by handling cross-sentence dependencies in components outside the decoder, e. g., by decoding in two passes (Le Nagard and Koehn, 2010; Xiao et al., 2011; Ture et al., 2012) or by using a special decoder driver module to annotate the decoder's input and recover the required information from its output (Hardmeier and Federico, 2010; Gong et al., 2012b). More recently, we have presented a decoding algorithm (Hardmeier et al., 2012) and a decoder (Hardmeier et al., 2013a) based on local search that permit the inclusion of cross-sentence feature functions directly into the decoding process, opening up new ways to design discourse-wide models. The integration of document-level features into the SMT decoding process is a central topic of this thesis and will be studied in Chapters 3 and 4.

2.5 Discourse-Aware MT Evaluation

A recurring issue in all discourse-related MT work is the problem of evaluation. The most popular automatic MT evaluation measure, BLEU (Papineni et al., 2002), calculates scores by measuring the overlap of low-order *n*-grams (usually up to 4-grams) between the output of the MT system and one or more reference translations. This score is insensitive to textual patterns that extend beyond the size of the *n*-grams, and it favours systems relying on strong *n*-gram models over other types of MT systems (Callison-Burch et al., 2006). It has been pointed out by various authors (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2011; Meyer et al., 2012) that this evaluation measure may not be adequate to guide research on specific discourse-related problems, and more targeted evaluation scores have been devised for the translation of pronominal anaphora (Hardmeier and Federico, 2010) and discourse connectives (Meyer et al., 2012; Hajlaoui and Popescu-Belis, 2012, 2013).

There has also been some effort to exploit discourse information to improve the evaluation of MT in general, independently of specific features in the MT systems tested. Giménez et al. (2010) propose an MT evaluation metric based on Discourse Representation Theory (Kamp and Reyle, 1993), which takes into account features like coreference relations and discourse relations to assess the quality of MT output. Unfortunately, their metric does not have a higher correlation with human quality judgements than standard sentence-level MT evaluation metrics in the MetricsMATR shared task (Callison-Burch et al., 2010). However, in more recent work, a metric using tree kernels (Collins and Duffy, 2002) over sentence-level discourse trees conforming to Rhetorical Structure Theory (Mann and Thompson, 1988) is shown to achieve a correlation approaching that of BLEU, and surpassing the current state of the art when combined with other metrics (Guzmán et al., 2014; Joty et al., 2014).

Wong et al. (2011) and Wong and Kit (2012) propose extending sentence-level evaluation metrics such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006) or METEOR (Banerjee and Lavie, 2005) with a component to measure lexical cohesion. For this purpose, they use measures of word repetition in the text, after applying either just stemming or semantic relatedness clustering according to similarity in WordNet (Fellbaum, 1998). They claim that there is a positive correlation between their lexical cohesion scores and human quality judgements, and that they can improve the correlation of BLEU and TER, but not METEOR, by combining them with the cohesion scores. In finding a positive correlation between lexical cohesion as measured by word repetition in MT output and human quality judgements, their results seem to be inconsistent with those of Carpuat and Simard (2012) discussed above, a discrepancy that should be investigated further to pin down the role of lexical cohesion in MT quality.

2.6 Conclusion

After an initial period during which SMT research, with very few exceptions (Marcu et al., 2000), was almost entirely uninterested in discourse-level processing, discourse-level and document-level aspects of translations have recently gained quite substantial attention. In a number of corpus studies, important challenges have been identified by studying such phenomena as word disambiguation (Carpuat, 2009; Carpuat and Simard, 2012; Ture et al., 2012), lexical cohesion (Voigt and Jurafsky, 2012; Guillou, 2013; Beigman Klebanov and Flor, 2013), pronominal anaphora (Hardmeier and Federico, 2010; Scherrer et al., 2011) or discourse connectives (Cartoni et al., 2011; Meyer et al., 2011b). Other discourse problems such as tense and aspect marking on verbs (Gong et al., 2012a,b; Meyer et al., 2013) and NP definiteness (Tsvetkov et al., 2013) have been studied more experimentally. All of these were shown to be highly relevant to translation quality, but in most cases it has been difficult to obtain noticeable improvements in BLEU scores or other empirical measures of MT quality. The most sizeable gains reported in the literature are for translation between English and Chinese (e. g., Gong et al., 2012a, with a tense model or Xiong et al., 2013a, with a lexical cohesion model). This may indicate that it is easier to achieve improvements if the distance between the languages is greater because it is more difficult for a baseline system to transfer information between dissimilar languages without the help of explicit models.

The most important reason for the limited success of existing discourse models for SMT is certainly that the underlying processes are not sufficiently understood for the creation of accurate models. The statistical approach to MT, which avoids all commitment to specific linguistic theories for the benefit of corpus-based pattern matching techniques, has been tremendously

successful, but as we begin to feel the limitations of the simple assumptions made in early SMT research, it becomes more and more difficult to extend the models without theoretical guidance. We hope that the research activities now begun sooner or later lead to an improved understanding of how different discourse processes affect translation that will, in turn, enable the development of better models.

Another serious problem is how to evaluate SMT system in a way that places due weight on discourse aspects. Just as progress in MT research in general was difficult to evaluate before the appearance of generally accepted automatic metrics such as BLEU (Papineni et al., 2002), the shortcomings of these automatic metrics when it comes to discourse make it difficult to assess progress in text-level MT. Complaints about the insensitivity of BLEU to discourse-level phenomena, even in cases where manual evaluation does find an improvement in the MT output, are common in the literature (e. g., Meyer et al., 2012; Taira et al., 2012; Novák et al., 2013a). While the final evaluation of an MT system can generally be done manually, the lack of good automatic evaluation metrics capturing discourse properties deprives discourse-enabled SMT systems from the possibility of automatically optimising model parameters toward translation quality. In sentence-level SMT, this is now a standard procedure that often results in significant improvements (Och, 2003).

In sum, there remains much to be done in the field of discourse-level SMT, even though there is considerably more research activity now than just a few years ago. In the remainder of this thesis, we try to make a contribution to two principal problems. In the first part, we investigate the interaction between discourse-level models and the decoding process in SMT and present a framework for document-level decoding that serves as a basis for further experimentation. In the second part, we investigate pronominal anaphora and the difficulties it poses for SMT.

Part I:
Algorithms for Document-Level SMT

3. Discourse-Level Processing with Sentence-Level Tools

In this chapter, we discuss the limitations of sentence-level SMT and some ways to overcome them while still using the same tools. First, we explain the principles of phrase-based SMT, the framework of all our experiments, and study the stack decoding algorithm, the most popular decoding algorithm for phrase-based SMT. We show how the stack decoder exploits model locality to increase decoding performance and why it is difficult to use document-level features in combination with this algorithm. Then, we examine three workarounds for the limitations of this algorithm and discuss their trade-offs and constraints, drawbacks and advantages.

3.1 An Overview of Phrase-Based SMT

There are a number of competing approaches to SMT, which differ in the way they decompose the input sentence and transfer its individual components into the target languages. Some of the most influential are phrase-based SMT (Koehn et al., 2003), hierarchical SMT (Chiang, 2007) and *n*-gram-based SMT (Mariño et al., 2006). All of these approaches model translation at the sentence level, and they have similar limitations when it comes to handling discourse phenomena. In this thesis, we concentrate on phrase-based SMT, and we shall not consider the other approaches any further. However, we expect all of them to present similar challenges, and we imagine that the considerations and solutions we propose are applicable to all forms of SMT, even though the implementation details are liable to vary. In this section, we give a brief overview of the aspects of phrase-based SMT that are relevant to our work. For a more detailed introduction, the reader is referred to the SMT textbook by Koehn (2010).

In the translation model of phrase-based SMT (Fig. 3.1), the input sentence is segmented into a sequence of non-overlapping word sequences (upper line)

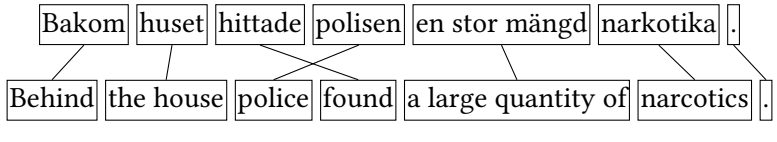

Figure 3.1. Sentence translation in phrase-based SMT

that are called *phrases*, even though they have little to do with phrases in the linguistic sense of the word. Each of the source language phrases in the input is mapped into a corresponding target language phrase (lower line). To account for differences in word order between the languages, the output phrases can be generated in an order that differs from that of their corresponding input phrases, or *reordered*. Given a realistic translation model, this procedure can generate an immense number of different hypotheses for an input sentence. Each hypothesis is then assigned a score by the model, and the goal is to find the translation that maximises the model score.

Modelling the quality of a hypothesis is difficult. It is easier to model different aspects that contribute to translation quality individually and combine these partial models into an overall score. By doing so, we can make different independence assumptions tailored to the structure of the partial models. The overall model score $f(s,t)$ of a target language output sentence t translating a given source language input sentence s is then computed as a linear combination of partial model scores, or feature functions, $h_k(s,t)$:

$$f(s,t) = \sum_k \lambda_k h_k(s,t) \tag{3.1}$$

Usually, the weights λ_k of the partial models are optimised discriminatively to maximise some automatic translation quality metric like BLEU (Papineni et al., 2002) with an optimisation technique such as MERT (Och, 2003), PRO (Hopkins and May, 2011) or MIRA (Chiang, 2012).

Unlike some other subfields of NLP such as syntactic parsing, where a similar model decomposition is used almost exclusively with binary feature functions indicating the presence or absence of a particular feature in the hypothesis, in SMT it is common to view Eq. 3.1 as a log-linear model (Berger et al., 1996), following Och and Ney (2002), and to use feature functions that represent log-transformed probability estimates. This has both historical and practical reasons. Early work on SMT (Brown et al., 1990, 1993) was strongly influenced by the standard methods in automatic speech recognition (Jelinek, 1976) and adopted the noisy channel model (Shannon, 1948) as its fundamental model. The noisy channel model corresponds to a log-linear model with uniform weights. The fact that reliable discriminative weight estimation for a large number of features in SMT has long been a difficult problem is an additional reason for preferring models with few, but informative features.

In principle, the partial models $h_k(s,t)$ can capture arbitrary features considered relevant to translation quality. There is a small set of models that are present in virtually any phrase-based SMT system and that are considered essential to achieve state-of-the-art performance. Usually, all phrase-based SMT systems will contain at least some variant of the following three models:

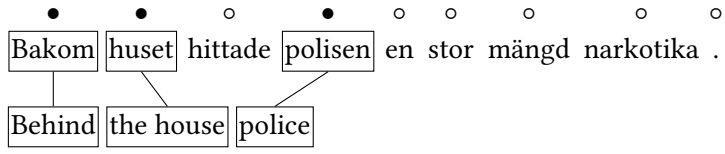

Figure 3.2. Stack decoding progress after translating 3 phrase pairs

- *Phrase translation model:* The phrase translation model assigns a probability score to the translation of a single SMT phrase in the source language to a given target language equivalent. It does not consider any context beyond the phrase boundaries.
- *Language model:* The language model is an *n*-gram model that assigns a probability to a target language word given a history of a bounded number of target language words to its left. It does not look at the input at all, and it only considers a limited number of context words for any given word.
- *Distortion model:* The distortion model assigns a probability to the order of the phrases in the output. In its basic form, it simply penalises differences in phrase order between the input and the output without looking at any further context or even at the words inside the phrases.

For decoding, it is important to notice that the use of context in these models is extremely limited. The translation model does not consider any phrase context at all. The basic distortion model only depends on the positions of the input words translated by the current and the immediately preceding phrase, and the language model depends on a bounded number of context words. This sparse and highly structured dependency configuration has been exploited to enable efficient decoding through dynamic programming.

3.2 The Stack Decoding Algorithm

The *de facto* standard algorithm for decoding phrase-based SMT models is a dynamic programming (DP) beam search algorithm commonly called *stack decoding* (Koehn et al., 2003). The stack decoding algorithm constructs a translation step by step by starting with an empty translation and adding words to it in target language word order while keeping track of which source language words are already covered by a phrase pair in the current translation hypothesis. At each step, the algorithm considers possible translations for input positions that are not yet covered and extends the state with another phrase pair until the entire input is covered. Figure 3.2 shows an example sentence after processing three phrase pairs. The top row indicates which words are covered. Next, the decoder will choose a new input phrase that covers one

or more of the uncovered input words and translate it into a phrase that will be appended to the output after the word "police".

In the stack decoder, incomplete hypotheses are grouped in *stacks* according to the number of input words they cover. Stacks are generic collections, unrelated to the last-in first-out data structure of the same name. The stacks are processed in order of ascending coverage count, beginning with the zero-coverage stack containing only the empty hypothesis and terminating with the final stack containing hypotheses that cover the entire input. Hypotheses on the individual stacks are expanded in order of descending score, and after processing a given number of items on each stack, the remaining hypotheses are ignored, or *pruned*. Because of pruning, stack decoding is a beam search algorithm.

The efficiency of stack decoding is greatly increased by a dynamic programming technique called *hypothesis recombination* (Och et al., 2001) that exploits the locality of the SMT models. Most of the complexity of a decoding algorithm is due to the fact that previously generated hypotheses must be processed over and over again whenever the scores are updated to add a new element. If dependencies are unrestricted, adding a new element may have the effect that a hypothesis which previously seemed suboptimal suddenly becomes best because it matches the new element better than any of the other hypotheses. This is why a large number of hypotheses must be stored and reexamined at each expansion step. However, since none of the basic models considers more than a few words of target language context, the dependencies of a new decision are very restricted in reality. Assuming a trigram language model, which considers a history of two words, all hypotheses that coincide in the last two words form an equivalence class from the point of view of future decisions. For each of these classes, only the best hypothesis need be retained; all others can be discarded without further ado because there is no way in which they can lead to the best overall translation. We say that they are *recombined* with the best hypothesis of their equivalence class. The beneficial effect of recombination is that it allows the decoder to explore a much larger part of the search space with the same stack size.

Consider now a situation in which one of the models has dependencies whose range substantially exceeds the history size of the n-gram model, as will usually be the case for the discourse phenomena that we are interested in. If the dependencies are long, but do not cross sentence boundaries, the stack decoding algorithm can still accommodate them. However, while the decoder generates the output between the two elements involved in the dependency, recombination will effectively be inhibited. As a result, the search space becomes much larger, and, assuming the stacks are pruned to the same size, the probability of making a search error will increase greatly. If the long-range dependencies cross sentence boundaries, the only way to handle this in the stack decoding algorithm is by suppressing the sentence boundaries and decoding the whole document, or a sufficiently large part of it to include all

the relevant dependencies, as if it were a single sentence. In this case, recombination will be inhibited almost completely, and the search space explosion described above will be exacerbated. For this reason, it has been necessary to find other ways to handle long-range dependencies in SMT decoding.

3.3 Two-Pass Decoding

Even though it is difficult to handle long-range dependencies in an SMT stack decoder, especially if they cross sentence boundaries, it is possible to use an unmodified sentence-level decoder to process certain discourse-level dependencies if decoding is carried out in two passes. This is the approach adopted by Le Nagard and Koehn (2010) and subsequently Guillou (2011, 2012) for their experiments with pronominal anaphora. It is also used to encourage translation consistency by Xiao et al. (2011) and Ture et al. (2012).

A model of pronominal anaphora must account for the agreement relation between anaphoric pronouns and the noun phrases they refer to (see Chapter 6 for an extended discussion). To transfer this relation into the target language, agreement must be ensured between the translation of the antecedent noun phrase and the translation of the anaphoric pronoun. Since both translations are generated by the SMT system, this implies modelling long-range target side dependencies, potentially across sentence boundaries. Le Nagard and Koehn (2010) address this problem by translating documents in two steps. First, they generate a translation from English into French with a normal SMT system without any knowledge of discourse or pronouns. Anaphoric links are resolved externally with a separate anaphora resolution system. When the first-pass translation is finished, the translations of the antecedents are recovered from the output, and the system looks up their gender. For all instances of the pronouns *it* and *they* identified as anaphoric by the anaphora resolution system, the gender of the translation is then marked on the input token, creating synthetic tokens such as `it-masculine`, and the text is translated again with an SMT system trained on this type of data.

This decoding approach is simple and has the advantage that it does not require any modifications to the existing software. Its drawback is mainly that the two-step procedure enforces categorical, hard decisions that make it difficult to create a coherent model of the problem as a whole. In particular, in the anaphora translation approach described above, all antecedent translations get fixed after the first translation step, and the system manipulates the anaphoric pronouns to encourage agreement. Formally, however, there is no guarantee that the second-pass translation step will select the same translations for the antecedent, so it is perfectly possible that the system translates the antecedent differently in the second pass and then enforces agreement with a purely fictitious antecedent translation that does not correspond to the final translation.

In the pronoun translation experiments published in the literature, this effect seems to be very small in practice. Guillou (2012), whose experiments on English–Czech are closely similar to the work on English–French described by Le Nagard and Koehn (2010), remarks that only 3 out of 458 antecedents were translated differently by her second-stage system. No corresponding figures are available for the original system by Le Nagard and Koehn (2010).

Guillou (2012) highlights that she takes extra care at training time to minimise the differences between her first- and second-stage system by making sure both systems are trained on exactly the same corpora and word alignments. The need to do this can make the training process fragile, but if it is carefully ensured, then the two systems can reasonably be expected to produce very similar output. Differences are most likely to occur when an antecedent and an anaphoric pronoun (referring to this or a different antecedent) occur close together in the text. In such cases, the influence of the n-gram model may trigger a different translation for the antecedent when the pronoun is translated differently in the second pass. Even if this kind of interference is rare with a simple pronoun model, it is much more likely to happen if more discourse-level models are incorporated into the same system using this approach.

Another limitation of the two-pass decoding approach is the directionality of its dependencies. Necessarily, with this method the overall model divides the relevant variables into two sets. One set (the antecedent translations) is fixed unconditionally in the first decoding pass. The other set (the pronoun translations) is assigned to in the second decoding pass with the possibility of conditioning on the values of the variables in the first set. The variables do not get optimised jointly, so there is no way in which the values of the variables in the second set can influence the choices made for the first set. In the case of pronominal anaphora, this is arguably the right way to model the phenomenon: Pronouns should agree with their nominal antecedents, but it is at least doubtful if the choice of a particular pronoun should ever induce a subsequent choice of a compatible antecedent noun phrase. However, this is not true of all kinds of discourse models. If the goal is to model, e. g., text cohesion by encouraging lexical consistency, it may well be advisable to optimise over the whole text jointly and combine information from different parts of the text rather than selecting the translation of the first word unconditionally and conditioning the rest of the text on this choice.

3.4 Sentence-to-Sentence Information Propagation

If the cross-sentence dependencies of a model form a directed acyclic graph, then it can be decoded with sentence-level tools without requiring two-pass decoding. This type of dependency configuration can reasonably be posited for models of pronominal anaphora. It is fairly safe to assume that cross-

sentence anaphoric links always introduce a dependency of an element (a pronoun) in a later sentence on an element (an antecedent) in an earlier sentence. The reverse situation, cross-sentence pronominal cataphora, is not impossible, but very uncommon in almost all text genres that are candidates for machine translation, so it can be neglected without great risk for translation quality, ensuring that all dependencies can be resolved in document order and no cycles occur.

The key to translating with cross-sentence dependencies is to decode each sentence individually instead of feeding the document to the decoder as a single batch. After each sentence has been translated, the information that is needed for translating later sentences can be extracted and fed into the decoder when it is time to do so. In the following paragraphs, we describe an approach to the integration of pronominal anaphora into an SMT system from our own work (Hardmeier and Federico, 2010). Gong et al. (2012b) use a similar procedure for decoding with a cross-sentence verb tense model.

Our system has two main components, a decoder driver, which encapsulates the sentence-based Moses decoder (Koehn et al., 2007) and propagates information between sentences, and a word dependency model, which injects information from previous sentences into the actual search process and handles sentence-internal coreference links. The word dependency model will be discussed in more detail in Chapter 7.

Figure 3.3 illustrates the workings of the decoder driver. Before the decoder is run, a sentence dependency graph (top right) is constructed based on the output of a separate coreference resolution system, BART (Versley et al., 2008). At the cross-sentence level, we only use anaphoric links. If there happen to be any cataphoric links, they are disregarded to guarantee that the sentence dependency graph is acyclic. Each sentence can contain pronominal mentions that refer to a preceding sentence (backward dependencies, marked *r*) as well as antecedent mentions that are referred to later (forward dependencies, marked *a*). The figure shows the state after translating sentences 1 and 2. Sentences that have no backward dependencies, such as sentences 1 and 2 in the example, and sentences whose backward dependencies have already been resolved, such as sentences 3 and 5, are put on a queue that feeds the decoder. After decoding, the translations of the antecedent mentions are recovered from the decoder output with the help of the phrase alignments produced by the decoder and the word alignments stored in the SMT phrase table. The decoder driver extracts the words aligned to what has been identified as the syntactic head of the antecedent mention and makes them available to the referring sentences by encoding them in the decoder input as described in the following section. Whenever all backward dependencies of a sentence are satisfied, the sentence is put on the queue.

The implementation described here makes it possible to feed a large number of decoder processes in a multi-threaded setup. The decoder input queue is realised as a priority queue ordered by the number of forward dependen-

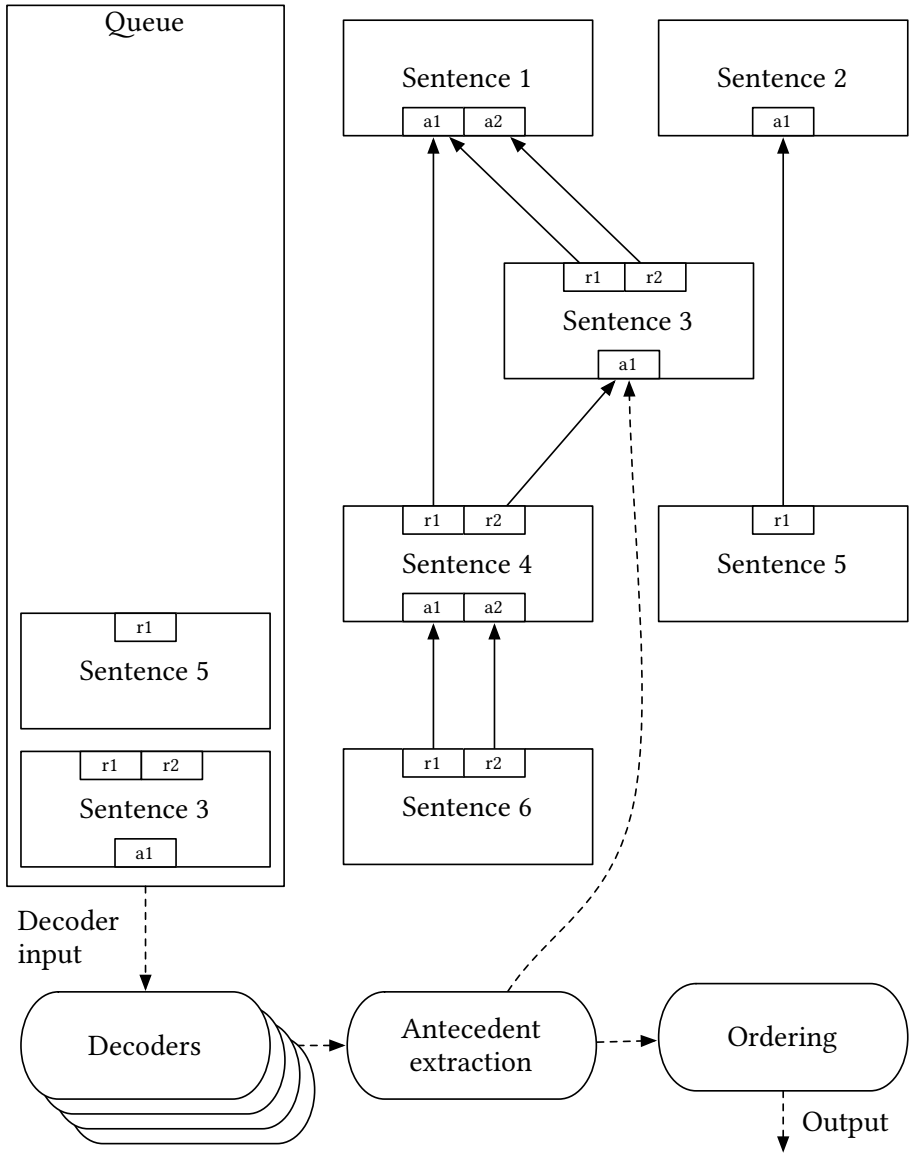

Figure 3.3. Decoder driver for sentence-to-sentence information propagation

cies of the sentences in order to resolve as many dependencies as possible as early as possible and thus increase the throughput of the system. Since the sentences are not processed in order, a final ordering step restores the original document order. For a slightly less complex setup, the dependency graph and the decoder input queue can be dispensed with, and the sentences can simply be processed in document order to ensure that the information from earlier sentences is available when it is needed.

The main advantage of the information propagation approach over the two-pass decoding procedure is that a single decoding pass is sufficient. This makes the approach slightly more efficient, but it is also attractive theoretically because it eliminates the potential discrepancies between the first- and the second-pass translation. In terms of dependency directionality, the constraints are the same. The information propagation approach requires the cross-sentence dependencies to form a directed acyclic graph, and translation decisions get fixed greedily as this graph is traversed with no opportunity for joint optimisation. The granularity of the dependency graph is at the sentence level; unlike two-pass decoding, the information propagation approach does not deal with sentence-internal dependencies. For a pronominal anaphora model, this is a problem because both intrasentential and intersentential anaphoric links are very frequent in corpus data, at least in the newswire genre (McEnery et al., 1997). In our work (Hardmeier and Federico, 2010), sentence-internal links are handled by the word dependency model in the decoder (see Chapter 7).

To sum up, information propagation is a fast and reliable approach for integrating discourse-level models into SMT if the dependency structure of all of these models mainly consists of cross-sentence links and complies with the constraints on the dependency graph imposed by the decoding procedure. In practice, all dependencies in all cross-sentence models must be directed and point in the same direction. If there are many sentence-internal dependencies, however, this approach will not help, and the usual constraints and limitations of standard stack decoding apply.

3.5 Document-Level Optimisation by Output Rescoring

One way to use models with unlimited dependencies on other sentences in combination with sentence-level SMT tools is to let the sentence-level system produce a variety of different output proposals and then perform a second search pass with the long-range models over the output variants suggested by the first-pass system only. This is the approach chosen, e. g., by Gong et al. (2011b) for integrating topic models into their SMT system. The search space of the second-pass rescoring step can be given either as an n-best list or in more compact form as a lattice representation of the part of the search space explored by the first pass decoder.

The advantage of this method is that it does not impose any restrictions at all on the models of the second-pass search, or on the number, type or orientation of any of the dependencies involved. It is possible to treat sentence-internal and cross-sentence dependencies in a uniform way. Moreover, the dependencies need not be oriented at all; if the search algorithm used for the second pass permits it, translations throughout the document can be optimised jointly and mutually influence each other. Since the search space of the second-pass search is relatively small, all this can be done efficiently.

The small size of the second-pass search space, which enables efficient search, is at the same time the main disadvantage of the rescoring approach. The size of the search space of phrase-based SMT is roughly exponential in the sentence length (Koehn, 2010, 161). By contrast, the number of complete hypotheses output by the stack decoding algorithm is bounded by a constant, the stack size. Therefore, the rescoring pass only gets to see an almost negligibly small subset of the search space. It is true that the construction of this subset with the stack decoding algorithm gives rise to hope that it may include some of the overall best translations, but since the first-pass decoder has no knowledge about the models to be included in the second pass, there is no formal guarantee that this is true even approximately under the second-pass models.

3.6 Conclusion

The stack decoding algorithm for phrase-based SMT cannot handle cross-sentence dependencies, and much of its efficiency is due to the fact that even sentence-internal dependencies are assumed to have very short ranges. Nevertheless, there are a number of possibilities to deal with discourse-level structure even in this framework. They all have differents strengths and weaknesses. Two-pass decoding and sentence-to-sentence propagation are similar. The former is a bit simpler and can potentially handle intrasentential dependencies, but there is a risk of inconsistencies, and the interaction between the decoder and model is difficult to analyse and understand. Also, the modelling possibilities are limited to what can be achieved by manipulating the translation model, unless specific models are implemented in the decoder, in which case the method loses its appealing simplicity. Both approaches require directed dependencies and do not support joint optimisation over the entire document. The *n*-best reranking method, by contrast, is unaffected by most of these limitations, but it can only access an exponentially small part of the entire search space. As a result, it is only suitable if there is reason to suppose that the best translation under the final model is already among the top candidates under the model the *n*-best lists are created with.

All of these techniques are most useful, and have been used almost exclusively, to integrate single models capturing specific features into the de-

coding process. With a greater number of cross-sentence features, or if the cross-sentence features have complex dependencies, they quickly become cumbersome and difficult to maintain. In the next chapter, we describe how discourse-level models can be fully integrated into SMT decoding. Like any other, our new approach has both advantages and drawbacks. Compared to the methods described in this chapter, however, it has a rather different profile, which makes it particularly interesting for large-scale experimentation with discourse models.

4. Document-Level Decoding with Local Search

In the previous chapter we studied different manners of handling document-level features with the standard tools of sentence-based SMT. We found that these approaches are limited in various ways and impose restrictions on the dependency configuration of the feature models or on the search space that can be explored. One of the goals of our work is to provide a framework for experimentation with discourse-level features in SMT that is as flexible as possible. It should be possible to experiment with different dependency configurations and restrictions to find out what setup best meets the needs of the modelling task. As far as possible, these constraints should not be imposed as a necessity by the decoding algorithm.

In this chapter, we present an approach to phrase-based SMT decoding where document-level features are completely integrated into the decoder (Hardmeier et al., 2012). We have released a software implementation of this approach, the Docent decoder, to the public (Hardmeier et al., 2013a). In order to escape the constraints of dynamic programming beam search, we abandon the stack decoding algorithm. Instead, we use a local search algorithm whose internal state consists of a complete translation of an entire document. This ensures that both the complete input document and a complete translation hypothesis are available whenever a score must be computed, so there are no restrictions placed on the dependencies of the feature models. Moreover, unlike a rescoring solution, our decoder has access to the entire search space of phrase-based SMT at least in principle, even though the vastness of the search space and the presence of local score maxima make search difficult. However, we show that our approach has reasonable performance in practice, and that it can be initialised with standard stack decoding to increase the chances of finding a good local maximum.

4.1 A Formal Model of Phrase-Based SMT

The phrase-based SMT model implemented by our decoder is exactly equivalent to the basic model of phrase-based SMT (Koehn et al., 2003), but it is formalised in a way that matches the properties of our decoding algorithm. The hypothesis space of our method is the same as that of sentence-level phrase-based SMT. In particular, we assume that the input is segmented into a number of sentences. The decoder emits exactly one output sentence for

each input sentence, and there is no mechanism to move information from one sentence into another. This assumption makes the decoder more compatible with existing SMT software and evaluation methods. Strictly speaking, however, it does not restrict its capabilities, since the entire document could always be presented to the decoder as a single "sentence".

Our decoder is based on local search, so its state at any time is a representation of a complete translation of the entire document. We decompose the state of a document into the state of its sentences, and we define the overall state S as a sequence of sentence states:

$$S = S_1 S_2 \dots S_N, \tag{4.1}$$

where N is the number of sentences.

Let i be the number of a sentence and m_i the number of input tokens of this sentence, p and q (with $1 \leq p \leq q \leq m_i$) be positions in the input sentence and $[p;q]$ denote the set of positions from p up to and including q. We say that $[p;q]$ precedes $[p';q']$, or $[p;q] \prec [p';q']$, if $q < p'$. Let $\Phi_i([p;q])$ be the set of translations for the source phrase covering exactly the positions $[p;q]$ in the input sentence i, as given by the phrase table. We call $A = \langle [p;q], \phi \rangle$ an *anchored phrase pair* with coverage $C(A) = [p;q]$ if $\phi \in \Phi_i([p;q])$ is a target phrase translating the source words at positions $[p;q]$. Then a sequence of n_i anchored phrase pairs

$$S_i = A_1 A_2 \dots A_{n_i} \tag{4.2}$$

is a valid sentence state for sentence i if the following two conditions hold:

1. The coverage sets $C(A_j)$ for j in $1, \dots, n_i$ are mutually disjoint, and
2. the anchored phrase pairs jointly cover the complete input sentence, or

$$\bigcup_{j=1}^{n_i} C(A_j) = [1; m_i]. \tag{4.3}$$

The MT output corresponding to a state is generated by iterating over the anchored phrase pairs in the order in which they occur in the state and reading off the target phrases ϕ of each anchored phrase pair.

Let $f(S)$ be a scoring function mapping a state S to a real number. As usual in SMT, it is assumed that the scoring function can be decomposed into a linear combination of K feature functions $h_k(S)$, each with a constant weight λ_k, so

$$f(S) = \sum_{k=1}^{K} \lambda_k h_k(S). \tag{4.4}$$

The decoder searches for the state $\hat{S}$ with maximal score, such that

$$\hat{S} = \arg\max_S f(S). \tag{4.5}$$

As a baseline, we implement a set of elementary feature functions compatible with the core features of the popular Moses SMT system (Koehn et al., 2007). All of these work on the sentence level, so a document-level decoder has no advantage if no discourse-level features are added. However, having this set of baseline feature functions is essential as a starting point for further development. In particular, our decoder has the following sentence-level feature functions:

1. Phrase translation scores including forward and backward conditional probabilities and lexical weights (Koehn et al., 2003),
2. n-gram language model scores implemented with the KenLM toolkit (Heafield, 2011),
3. a word penalty score,
4. a phrase penalty score,
5. a distortion model with geometric decay (Koehn et al., 2003), and
6. a feature indicating the number of times a given distortion limit is exceeded in the current state.

The baseline features are computed at the sentence level, and the document score is just the sum over all sentence scores. In our experiments, the last feature is used with very large negative fixed weight in order to limit the gaps between the coverage sets of adjacent anchored phrase pairs to a maximum value. In DP search, the distortion limit is enforced directly by the search algorithm to limit complexity. In our decoder, however, this restriction is not required, so we add it among the scoring models. In principle, its weight could be determined automatically during feature weight optimisation (Stymne et al., 2013b).

4.2 The Local Search Decoding Algorithm

The decoding algorithm we use (Algorithm 1) is very simple. It starts with a given initial document state. In the main loop, which extends from line 3 to line 12, it generates a successor state S' for the current state S by calling the function `Neighbour`, which non-deterministically applies one of the operations described in Section 4.4 to S. The score of the new state is compared to that of the previous one. If it meets a given acceptance criterion, S' becomes the current state, else search continues from the previous state S. The main loop is repeated until a maximum number of steps (*step limit*) is reached or until a maximum number of moves are rejected in a row (*rejection limit*).

For the experiments in this chapter, we use the *hill climbing acceptance criterion*, which simply accepts a new state if its score is higher than that of the current state. It is defined as

$$\texttt{Accept}(\alpha', \alpha) = \begin{cases} \textbf{true} & \text{if } \alpha' > \alpha \\ \textbf{false} & \text{otherwise.} \end{cases} \tag{4.6}$$

Algorithm 1 Decoding algorithm

Input: an initial document state S;
search parameters *maxsteps* and *maxrejected*
Output: a modified document state

```
1:  nsteps ← 0
2:  nrejected ← 0
3:  while nsteps < maxsteps and
    nrejected < maxrejected do
4:    S′ ← Neighbour(S)
5:    if Accept(f(S′), f(S)) then
6:      S ← S′
7:      nrejected ← 0
8:    else
9:      nrejected ← nrejected + 1
10:   end if
11:   nsteps ← nsteps + 1
12: end while
13: return S
```

The hill climbing criterion guarantees that the score never decreases in the course of decoding. However, it only permits state modifications that improve the score in a single step. Changes that require going through intermediate steps with lower scores, for instance to split up a phrase pair into smaller units before modifying a part of it, are impossible.

A notable difference between our algorithm and other hill climbing algorithms previously used for SMT decoding (Germann et al., 2004; Langlais et al., 2007; see Section 4.8) is its non-determinism. Earlier work on sentence-level decoding employed a *steepest ascent* strategy which amounts to enumerating the complete neighbourhood of the current state as defined by the state operations and selecting the next state to be the best state found in the neighbourhood of the current one. Enumerating all neighbours of a given state, costly as it is, has the advantage that it makes it easy to prove local optimality of a state by recognising that all possible successor states have lower scores. It can be rather inefficient, since at every step only one modification will be adopted; many of the modifications that are discarded will very likely be generated anew in the next iteration.

As we extend the decoder to the document level, the size of the neighbourhood that would have to be explored in this way increases considerably. Moreover, the inefficiency of the steepest ascent approach potentially increases as well. Very likely, a promising move in one sentence will remain promising after a modification has been applied to another sentence, even though this is not guaranteed to be true in the presence of document-level models. We therefore adopt a *first-choice hill climbing* strategy that non-

deterministically generates successor states and accepts the first one that meets the acceptance criterion. This frees us from the necessity of generating the full set of successors for each state. On the downside, if the full successor set is not known, it is no longer possible to prove local optimality of a state, so we are forced to use a different condition for halting the search. We use a combination of two limits: The *step limit* is a hard limit on the resources the user is willing to expend on the search problem. The value of the *rejection limit* determines how much of the neighbourhood is searched for better successors before a state is accepted as a solution; it is related to the probability that a state returned as a solution is in fact locally optimal.

It is also possible to combine Algorithm 1 with another acceptance criterion than that of Eq. 4.6. In particular, an acceptance criterion that sometimes accepts new states with lower scores than the current one may help the decoder to reach better states that are only accessible through a sequence of moves. In the Docent decoder, we also implement search by *simulated annealing* (Kirkpatrick et al., 1983) with the *Metropolis-Hastings acceptance criterion* (Metropolis et al., 1953; Hastings, 1970). This is a stochastic criterion defined as

$$\texttt{Accept}(\alpha', \alpha) = \begin{cases} \textbf{true} & \text{with probability } A(\alpha', \alpha; T) \\ \textbf{false} & \text{with probability } 1 - A(\alpha', \alpha; T) \end{cases} \tag{4.7}$$

with an acceptance probability satisfying

$$A(\alpha', \alpha; T) = \begin{cases} 1 & \text{if } \alpha' > \alpha \\ \exp\left(\dfrac{\alpha' - \alpha}{T}\right) & \text{otherwise.} \end{cases} \tag{4.8}$$

The temperature parameter T starts at a high value and is gradually reduced according to some *cooling schedule* as decoding progresses. As T approaches 0, the Metropolis-Hastings criterion in Eq. 4.7 becomes equal to the hill climbing criterion in Eq. 4.6, and indeed, the Docent decoder implements hill climbing as a special case of simulated annealing.

The asymptotic behaviour of simulated annealing search depends on the distribution of the transition probabilities from one state to the next. The transition probabilities are determined by the interaction of the proposal distribution embodied in the `Neighbour` function, which generates new states from the current one, and the acceptance distribution represented by the `Accept` function. In our system, the proposal distribution is controlled by the set of state operations described in Section 4.4 and their weights. If the transition probabilities satisfy a condition called *detailed balance*, then simulated annealing is guaranteed to converge to a global optimum asymptotically (Aarts et al., 1997). One way to meet this condition is to use the Metropolis-Hastings acceptance criterion in conjunction with a proposal distribution that guarantees that all states can be reached from all other states through

a sequence of operations with nonzero probabilities and is *symmetric*, meaning that for all pairs of states S and S', the probability of proposing state S' when in state S is equal to the probability of proposing state S when in state S' (Aarts et al., 1997, Theorem 3). Our current set of state operations does not satisfy the symmetry condition, so we cannot be sure that our simulated annealing procedure converges to an optimal solution even asymptotically.[1]

Empirically, the main difficulty with using simulated annealing instead of hill climbing for SMT decoding is that it is very easy for the decoder to wander off quickly to states with very bad scores from which it never finds its way back to better solutions. We have not analysed this behaviour in detail, but it seems likely that it is related to the irregularity of the proposal distribution mentioned in the previous paragraph and could be remedied by designing better proposal distributions. This, however, is a problem that we must leave to future work. Instead, we control the simulated annealing search process with some specific state operations that help the decoder return more easily to good states it has visited before. These operations are described at the end of Section 4.4.

4.3 State Initialisation

Before the local search decoding algorithm can be run, an initial state must be generated. The closer the initial state is to an optimum, the less work remains to be done for the algorithm. If the algorithm is to be self-contained, initialisation must be relatively uninformed and can only rely on some general prior assumptions about what might be a good initial guess. On the other hand, if optimal results are sought after, it pays off to invest some effort into a good starting point. One way to do this is to run DP search first.

For uninformed initialisation, we implement a very simple procedure based only on the observation that, at least when translating between the major European languages, it is usually a good guess to keep the word order of the output very similar to that of the input. We therefore create the initial state by selecting, for each sentence in the document, a random sequence of randomly segmented anchored phrase pairs covering the input sentence in monotonic order, that is, such that for all pairs of adjacent anchored phrase pairs A_j and A_{j+1}, we have that $C(A_j) \prec C(A_{j+1})$.

For initialisation with DP search, we first run the Moses decoder (Koehn et al., 2007) to generate an initial state. Then we extract the best output hypothesis from the Moses search graph and interpret it as a sequence of anchored phrase pairs. In Moses, we include a relaxed version of the models of the document-level decoding pass, omitting all models with document-level de-

[1]Technically, the conditions described are sufficient, but not necessary, for detailed balance. We do not expect detailed balance to obtain in our decoder, but we must defer a more rigorous analysis to the future.

pendencies. In the experiments of this thesis, we generally use a configuration as similar as possible to that of the document-level decoder with the same set of sentence-level models and the same feature weights.

4.4 State Operations

Given a document state S, the decoder uses a neighbourhood function called `Neighbour` to simulate a move in the state space. The neighbourhood function non-deterministically selects a type of state operation and a location in the document to apply it to and returns the resulting new state. In practice, operations are selected by drawing randomly from a categorical distribution with configurable, fixed parameters. To allow the decoder to explore the entire search space, it must be possible to alter the phrase segmentation of the input, the translations of the individual phrases as well as their output order. By selecting a set of operations geared towards these three aspects we can ensure that every possible document state can be reached from every other state in a sequence of moves.

Designing operations for state transitions in local search for phrase-based SMT is a problem that has been addressed in the literature (Langlais et al., 2007; Arun et al., 2010). Our decoder's first-choice hill climbing strategy never enumerates the full neighbourhood of a state. We therefore place less emphasis than previous work on defining a compact neighbourhood, but allow the decoder to make quite extensive changes to a state in a single step with a certain probability. Otherwise our operations are similar to those used by Arun et al. (2010).

All of our state operations except those described in Section 4.4.4 make changes to a single sentence only. Each time it is called, the `Neighbour` function selects a sentence in the document with a probability proportional to the number of input tokens in each sentence to ensure a fair distribution of the decoder's attention over the words in the document regardless of varying sentence lengths.

To simplify notations in the description of the individual state operations, we write

$$S_i \longrightarrow S_i' \tag{4.9}$$

to signify that a state operation, when presented with a document state as in Eq. 4.1 and acting on sentence i, returns a new document state of

$$S' = S_1 \ldots S_{i-1}\ S_i'\ S_{i+1} \ldots S_N. \tag{4.10}$$

Similarly,

$$S_i : A_j^{j+h-1} \longrightarrow \tilde{A}_1^{h'} \tag{4.11}$$

is equivalent to

$$S_i \longrightarrow A_1^{j-1}\ \tilde{A}_1^{h'}\ A_{j+h}^{n_i} \tag{4.12}$$

with

$$A_j^{j+h} \equiv A_j \ldots A_{j+h} \tag{4.13}$$

and indicates that the operation returns a state in which a sequence of h consecutive anchored phrase pairs has been replaced by another sequence of h' anchored phrase pairs.

4.4.1 Changing Phrase Translations

The `change-phrase-translation` operation replaces the translation of one single phrase with a random translation with the same coverage taken from the phrase table. Formally, the operation selects an anchored phrase pair A_j by drawing uniformly from the elements of S_i and then draws a new translation ϕ' uniformly from the set $\Phi_i(C(A_j))$. The new state is given by

$$S_i : A_j \longrightarrow \langle C(A_j), \phi' \rangle. \tag{4.14}$$

4.4.2 Changing Phrase Order

There are different useful ways to change the order of the output phrases. Our basic phrase order operation, used in all experiments described in this chapter, is called `swap-phrases`. It affects the output word order without changing the phrase translations. It exchanges two sequences of anchored phrase pairs of lengths l_1 and l_2, resulting in an output state of

$$S_i : A_j^{j+l_1+h+l_2-1} \longrightarrow A_{j+l_1+h}^{j+l_1+h+l_2-1}\ A_{j+l_1}^{j+l_1+h-1}\ A_j^{j+l_1-1} \tag{4.15}$$

The start location j is drawn uniformly from the eligible sentence positions; the swap range h and the lengths l_1 and l_2 come from geometric distributions with configurable decays.

Another reasonable option is the `move-phrases` operation, which moves a sequence of anchored phrase pairs either to the left or to the right without requiring any other phrase pairs to make the corresponding opposite movement. The resulting output states are

$$S_i : A_j^{j+h+l-1} \longrightarrow A_{j+l}^{j+h+l-1}\ A_j^{j+l-1} \tag{4.16}$$

for a right move and

$$S_i : A_j^{j+h+l-1} \longrightarrow A_{j+h}^{j+h+l-1}\ A_j^{j+h-1} \tag{4.17}$$

for a left move. The move direction is selected randomly, and the start location j, the jump distance h and the length l are determined in the same way as for the `swap-phrases` operation. Left and right moves are equivalent, but the effects of the parameters of the distributions of h and l are exchanged.

4.4.3 Resegmentation

The most complex operation is `resegment`, which allows the decoder to alter the segmentation of the source phrase. It takes a number of anchored phrase pairs that form a contiguous block both in the input and in the output and replaces them with a new set of phrase pairs covering the same span of the input sentence. Formally,

$$S_i : A_j^{j+h-1} \longrightarrow \tilde{A}_1^{h'} \tag{4.18}$$

such that

$$\bigcup_{k=j}^{j+h-1} C(A_k) = \bigcup_{k=1}^{h'} C(\tilde{A}_k) = [p;q] \tag{4.19}$$

for some p and q, where, for $k = 1,\ldots,h'$, we have that $\tilde{A}_k = \langle [p_k;q_k], \phi_k \rangle$, all coverage sets $[p_k;q_k]$ are mutually disjoint and each ϕ_k is randomly drawn from $\Phi_i([p_k;q_k])$. Regardless of the ordering of A_j^{j+h-1}, the `resegment` operation always generates a sequence of anchored phrase pairs in linear order, such that $C(\tilde{A}_k) \prec C(\tilde{A}_{k+1})$ for $k = 1,\ldots,h'-1$.

As for the other operations, j is generated uniformly and h is drawn from a geometric distribution with a decay parameter. The new segmentation is generated by extending the sequence of anchored phrase pairs with random elements starting at the next free position, proceeding from left to right until the whole range $[p;q]$ is covered.

4.4.4 Special Operations for Simulated Annealing

As discussed above (Section 4.2), combining the operations described so far with the Metropolis-Hastings acceptance criterion instead of pure hill climbing often leads the decoder astray, making it abandon promising hypotheses too easily and spend inordinate amounts of time on low-scoring parts of the search space. To reduce this risk, we introduce two operations that make simulated annealing behave more like hill climbing by frequently offering it short cuts back to good states.

The `restore-best` operation quite simply keeps track of the best state encountered during the current decoding run and offers it to the decoder again regardless of what the current state looks like. By its nature, it will always be accepted. The more frequently this operation is invoked, the more the search resembles hill climbing. If it is added to the proposal distribution with a relatively low probability, simulated annealing will have the opportunity to make excursions to lower-scoring states, but it will always be sent back to the original hill climbing path at some point unless it manages to find a better path in the meantime. Using this operation allows us to exploit some of the flexibility of simulated annealing whilst preserving the reliability of hill climbing.

The `crossover` operation bears some resemblance to the way a genetic search algorithm generates hypotheses. Like `restore-best`, it keeps track of the best state encountered. Instead of just going back to that state, it creates a new state which is a combination of the current state and the cached best state. For each sentence in the new state, it stochastically selects the corresponding sentence state from one of the two source states. The probability with which the better state is preferred is a parameter of the operation. This operation makes it possible to restore the safer choices of the previously best state for some sentences while allowing the current state arrived at by simulated annealing to retain some of its features.

When decoding with the hill climbing acceptance criterion, the current state is necessarily always the best state encountered so far, so these two operations would have no effect in the form described here. An operation similar to `crossover` could certainly be defined for the hill climbing case as well by selecting the second source state in some other way.

4.5 Efficiency Considerations

When implementing feature functions for the local search decoder, we have to exercise some care to avoid recomputing scores for the whole document at every iteration. To achieve this, the scores are computed completely only once, at the beginning of the decoding run. In subsequent iterations, the scoring functions are presented with the scores of the previous iteration and a list of modifications produced by the state operation, a set of tuples $\langle i, r, s, \tilde{A}_1^{h'} \rangle$, each indicating that the document should be modified as described by

$$S_i : A_r^s \longrightarrow \tilde{A}_1^{h'}. \tag{4.20}$$

If a feature function is decomposable in some way, as all the standard features developed under the constraints of DP search are, it can then update the state simply by subtracting and adding score components pertaining to the modified parts of the document. Feature functions have the possibility to store their own state information along with the document state to make sure the required information is available. Thus, the framework makes it possible to exploit decomposability for efficient scoring without imposing any particular decomposition on the features as DP beam search does.

To make scoring even more efficient, scores are computed in two passes: First, every feature function is asked to provide an upper bound on the score that will be obtained for the new state. For any feature function that represents a log-transformed probability, 0 is a trivial upper bound, but in many cases, it is possible to calculate much tighter upper bounds far more efficiently than computing the exact feature value, e. g., by removing just a small number of terms related to words that are affected by a proposed state change in a larger summation. If the upper bound fails to meet the acceptance criterion,

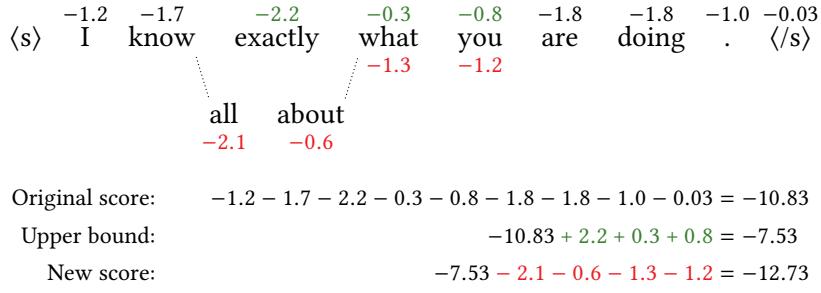

Figure 4.1. Two-pass LM score computation with a trigram LM

the new state is discarded right away; if not, the full score is computed and the acceptance criterion is tested again.

Among the basic models listed at the end of Section 4.1, this two-pass strategy is only used for the *n*-gram LM, which requires fairly expensive parameter lookups for scoring. The scores of all the other baseline models are fully computed during the first scoring pass. The *n*-gram model is more complex. Figure 4.1 illustrates how the LM implementation in the Docent decoder proceeds to compute first an upper bound, then an updated score as a word in the document state (*exactly*) is replaced by a sequence of two other words (*all about*). In its state information, the *n*-gram model keeps track of the LM score and LM library state for each word. The first scoring pass then identifies the words whose LM scores are affected by the current search step. This includes the words changed by the search operation as well as the words whose history is modified. In our implementation, the range of the history dependencies can be determined precisely by considering the "valid state length" information provided by the KenLM language modelling library (Heafield, 2011). In the first pass, the LM scores of the affected words are subtracted from the total score. The model only looks up the new LM scores for the affected words and updates the total score if the new search state passes the first acceptance check. This two-pass scoring approach allows us to avoid language model lookups altogether for states that will be rejected anyhow because of low scores from the other models, e. g., because the distortion limit is violated.

Model score updates become more complex and slower as the number of dependencies of a model increases. While our decoding algorithm does not impose any formal restrictions on the number or type of dependencies that can be handled, there will be practical limits beyond which decoding becomes unacceptably slow or the scoring code becomes very difficult to maintain. However, these limits are fairly independent of the types of dependencies handled by a model, which permits the exploration of more varied model types than those handled by DP search.

4.6 Experimental Results

In this section, we present the results of a series of experiments with our document decoder. The goal of our experiments is to demonstrate the behaviour of the decoder and characterise its response to changes in the fundamental search parameters. In all experiments presented in this chapter, we use the hill climbing acceptance criterion and the baseline set of sentence-level feature functions listed in Section 4.1. The search operations of the document decoder are `change-phrase-translation` with a weight of 0.8, `swap-phrases` with a weight of 0.1 and a swap distance decay of 0.5 and `resegment` with a weight of 0.1 and a resegmentation length decay of 0.1.

The SMT models for our experiments were created with a subset of the training data for the English–French shared task at the WMT 2011 workshop (Callison-Burch et al., 2011). The phrase table was trained on Europarl, News commentary and UN data. To reduce the training data to a manageable size, singleton phrase pairs were removed before the phrase scoring step. Significance-based filtering (Johnson et al., 2007) was applied to the resulting phrase table, and all phrase pairs not ranking among the top 20 per source phrase in terms of the conditional probability of the target phrase given the source phrase were discarded. The language model was a 5-gram model with Kneser-Ney smoothing trained on the monolingual News corpus with IRSTLM (Federico et al., 2008). Feature weights were trained with Minimum Error-Rate Training (MERT; Och, 2003) on the *news-test2008* development set using the DP beam search decoder and the MERT implementation of the Moses toolkit (Koehn et al., 2007). Experimental results are reported for the *newstest2009* test set, a corpus of 111 newswire documents totalling 2,525 sentences or 65,595 English input tokens.

4.6.1 Stability

An important difference between our decoder and the classical DP decoder as well as previous work in SMT decoding with local search is that our decoder is inherently non-deterministic. This implies that repeated runs of the decoder with the same search parameters, input and models will not, in general, find the same local maximum of the search space. The first empirical question we ask is therefore how different the results are under repeated runs. The results in this and the next section were obtained with the uninformed state initialisation described in Section 4.3, i. e., without running the DP beam search decoder.

Figure 4.2 shows the results of 7 decoder runs with the models described above, translating the *newstest2009* test set, with a step limit of $2^{27} \approx 1.3 \cdot 10^8$ and a rejection limit of 100,000. The x-axis of both plots shows the number of decoding steps on a logarithmic scale, so the number of steps is doubled between two adjacent points on the same curve. In the left plot, the y-axis

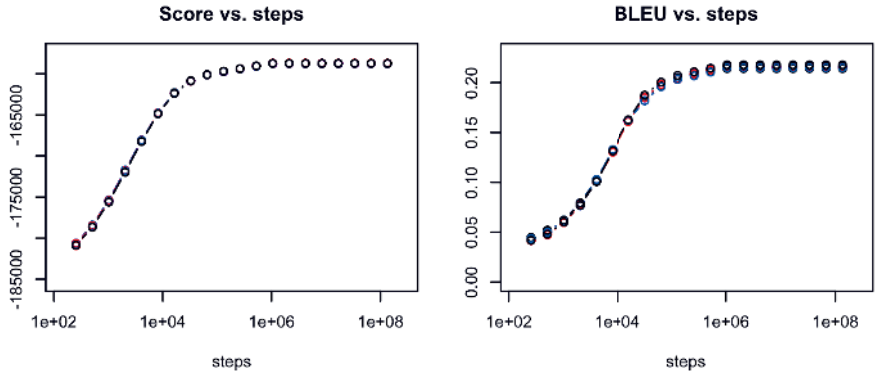

Figure 4.2. Score stability in repeated decoder runs

indicates the model score optimised by the decoder summed over all 2,525 sentences of the document. In the right plot, the case-sensitive BLEU score (Papineni et al., 2002) of the current decoder state against a reference translation is displayed.

As expected, the decoder achieves a considerable improvement of the initial state with diminishing returns as decoding continues. Between $2^8 = 256$ and $2^{14} = 16{,}384$ steps, the score increases at a roughly logarithmic pace, then the curve flattens out, which is partly due to the fact that decoding for some documents stops after the maximum number of rejections has been reached. The BLEU score curve shows a similar increase, from an initial score below 0.05 to a maximum of around 0.215. This is below the score of 0.2245 achieved by the stack decoder with the same models. The lower score is not surprising considering that our decoder approximates a more difficult search problem, from which a number of strong independence assumptions have been lifted, without, at the moment, having any stronger models at its disposal to exploit this additional freedom for better translation.

In terms of stability, there are no dramatic differences between the decoder runs. The small differences that exist are hardly discernible in the plots. The model scores at the end of the decoding run range between -158767.9 and -158716.9, a relative difference of only about 0.03 %. Final BLEU scores range from 0.2141 to 0.2163, an interval that is not negligible, but comparable to the variance observed when, e. g., feature weights from repeated MERT runs are used with one and the same SMT system. Note that these results were obtained with random state initialisation. With DP initialisation, score differences between repeated runs rarely exceed 0.02 absolute BLEU percentage points, but the improvement achievable with the baseline feature models is hardly any greater than this because the hypothesis found by the DP decoder is nearly optimal already.

Overall, we conclude that the decoding results of our algorithm are reasonably stable despite the non-determinism inherent in the procedure. In the

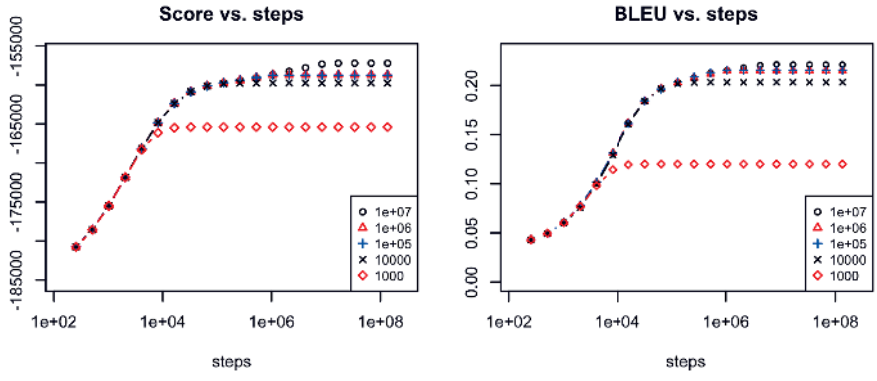

Figure 4.3. Search performance at different rejection limits

remaining experiments of this chapter, the evaluation scores reported are calculated as the mean of three runs for each experiment.

4.6.2 Search Algorithm Parameters

The hill climbing algorithm we use has two parameters which govern the trade-off between decoding time and the accuracy with which a local maximum is identified: The *step limit* stops the search process after a certain number of steps regardless of the search progress made or lack thereof. The *rejection limit* stops the search after a certain number of unsuccessful attempts to make a step, when continued search does not seem to be promising. In most of our experiments, we set the step limit to $2^{27} \approx 1.3 \cdot 10^8$ and the rejection limit to 10^5. In practice, decoding terminates by reaching the rejection limit for the vast majority of documents. We therefore examine the effect of different rejection limits on the learning curves. The results are shown in Fig. 4.3.

The results show that continued search does pay off to a certain extent. Indeed, the curve for rejection limit 10^7 seems to indicate that the model score increases steadily, albeit more slowly, even after the curve has started to flatten out at $2^{14} = 16{,}384$ steps. At a certain point, however, the probability of finding a good successor state drops rather sharply by about two orders of magnitude, as evidenced by the fact that a rejection limit of 10^6 does not give a large improvement over one of 10^5, while one of 10^7 does, so searching the state neighbourhoods very thoroughly gives a reward. The continued model score improvement also results in an increase in BLEU scores, and with an average BLEU score of 0.221 the system with rejection limit 10^7 is fairly close to the score of 0.2245 obtained by DP beam search.

Obviously, more exact search comes at a cost. In this case, it comes at a considerable cost, which is an explosion of the time required to decode the test set from 4 minutes at rejection limit 10^3 to 224 minutes at rejection limit 10^5

and 38 hours 45 minutes at limit 10^7. The DP decoder takes 31 minutes for the same task. We conclude that the rejection limit of 10^5 selected for our experiments, while technically suboptimal, realises a good trade-off between decoding time and accuracy.

4.7 Feature Weight Optimisation

As usual in SMT, our document-level decoder decomposes its objective function into a linear combination of partial models (Eq. 4.4). For the best possible translation quality, the feature weights λ_i should be optimised with a held-out development set. Fortunately, some of the weight optimisation methods from sentence-based SMT can be applied at the document level, too. In particular, MERT (Och, 2003) often works reasonably well for document-level weight tuning with only minor changes.[2]

MERT is an optimisation procedure which finds a set of feature weights directly optimising an automatic translation quality measure such as BLEU. It works with a representation of the search space as a list of translation hypotheses. In practice, therefore, the SMT search space is too large to be searched exhaustively. Instead, it is approximated with n-best lists. Since n-best lists only cover an exponentially small subset of the search space and are strongly biased, the resulting feature weights are not optimal for the entire space in general. To find good weights, MERT is run repeatedly. After each MERT run, the tuning set is translated again with the new feature weights to produce a new n-best list, which is then added to the list of the previous iteration before MERT is called again. This procedure is typically repeated until the list becomes stable and no new translations get added to the list in one iteration.

Adapting MERT to the document level requires two changes. The first concerns score computation. The data points considered by the MERT optimiser now represent complete documents instead of single sentences because no meaningful scores are available at the sentence level. Conceptually, this is a very simple change, but it has the effect that the number of data points for a given amount of tuning data becomes much lower. This may lead to reduced stability, but Stymne et al. (2013a) find that the amount of data in typical tuning sets is often sufficient to achieve useful results.

The second problem is the generation of n-best lists with a hill climbing decoder. Since the hill climbing algorithm never accepts downhill moves, the n-best output of this decoder will always consist of the last n accepted states. As a result of their construction with the state operations described above, these states will be very similar to each other, and the overall variety of the n-best list will be much smaller than that produced by a stack decoder. Con-

[2]The results on document-level feature weight optimisation presented in this section are joint work with Sara Stymne, Jörg Tiedemann and Joakim Nivre (Stymne et al., 2013a). The experiments were carried out by Sara Stymne.

sequently, the MERT optimiser will see an even smaller and more biased part of the search space, leading to bad feature weight estimates. The solution proposed by Stymne et al. (2013a) is to replace the n-best lists with more general n-lists obtained by sampling at regular intervals during the optimisation process. The optimal sampling conditions still need to be investigated more precisely.

4.8 Related Work

Even though DP beam search in the form of stack decoding (Koehn et al., 2003) has been the dominant approach to SMT decoding in recent years, methods based on local search have been explored at various times. For word-based SMT, greedy hill climbing techniques were advocated as a faster replacement for DP beam search (Germann et al., 2001; Germann, 2003; Germann et al., 2004), and a problem formulation specifically targeting word reordering with an efficient word reordering algorithm has been proposed (Eisner and Tromble, 2006).

A sentence-level local search decoder has been advanced as an alternative to the stack decoding algorithm also for phrase-based SMT (Langlais et al., 2007, 2008). That work anticipates many of the features found in our decoder, including the use of local search to refine an initial hypothesis produced by DP beam search. The possibility of using models that do not fit well into the DP paradigm is mentioned and illustrated with the example of a reversed n-gram language model, which the authors claim would be difficult to implement in a DP decoder. Similarly to the work by Germann et al. (2001), their decoder is deterministic and explores the entire neighbourhood of a state in order to identify the most promising step. Our main contribution with respect to the work by Langlais et al. (2007) is the introduction of the possibility of handling document-level models by lifting the assumption of sentence independence. As a consequence, enumerating the entire neighbourhood becomes too expensive, which is why we resort to a "first-choice" strategy that non-deterministically generates states and accepts the first one encountered that meets the acceptance criterion.

More recently, Gibbs sampling has been proposed as a way to generate samples from the posterior distribution of a phrase-based SMT decoder (Arun et al., 2009, 2010), a process that resembles local search in its use of a set of state-modifying operators to generate a sequence of decoder states. Where local search seeks for the best state attainable from a given initial state, Gibbs sampling produces a representative sample from the posterior. Like all work on SMT decoding that we know of, the Gibbs sampler presented by Arun et al. (2010) assumes independence of sentences and considers the complete neighbourhood of each state before taking a sample.

4.9 Conclusion

In this chapter, we have presented a document-level decoder for phrase-based SMT. The decoder (Hardmeier et al., 2012, 2013a) uses a local search approach, keeping a translation of the entire document as its internal state and continually generating new hypotheses by applying state-modifying operations to the current state. New states are accepted or rejected according to an acceptance criterion that deterministically or stochastically favours states with higher scores. Compared to the standard DP beam search algorithm, stack decoding, our approach has the advantage of admitting unrestricted dependency configurations for the feature models. On the downside, our algorithm explores a much larger search space than the stack decoder without profiting from the benefits of DP, so given models that are compatible with the constraints of DP, the risk of search errors is much increased. However, we have shown that our decoder on its own can generate translations whose BLEU score is only about one point lower than that of the translations found by Moses with the same models. Moreover, if we initialise the decoder with Moses output and use the hill climbing acceptance criterion, we know with certainty that only model error, not search errors, can make the final translations worse than those found by Moses.

Compared to the approaches to document-level SMT discussed in the previous chapter, our integrated document-level decoder has a number of advantages. The most important may well be its flexibility. While the sentence-based approaches all impose their specific restrictions on the models and make it difficult to experiment freely with discourse-level models, our decoder has no such inherent restrictions. It gives the feature models access to the entire document and permits joint optimisation of the feature functions over the complete document without constraining the directionality of the dependencies. It can accommodate any number of discourse-level features without additional complications. Its search algorithm is less efficient than DP beam search when it operates under the same constraints, but its performance does not suffer additionally from the presence of long-range dependencies. A sentence-based decoding procedure may be sufficient for some types of document-level models, and it may even be more efficient in some specific cases, but a document-level decoder provides an indispensable framework for unfettered experimentation with discourse features in SMT.

5. Case Studies in Document-Level SMT

In this chapter, we look at how we can apply the document-level decoding method of the previous chapter to control properties of the target language vocabulary of an SMT system. First, we consider the problem of lexical cohesion and terminological consistency in MT. We describe the results of a small corpus study and present a cross-sentence semantic language model based on a vector space representation. Then, we discuss how discourse models can be used to bias an SMT system towards certain types of vocabulary and show some results with document-level features to improve text readability.

5.1 Translating Consistently: Modelling Lexical Cohesion

Text cohesion, the property of linkedness in a text, is created not only by overt devices such as discourse markers or anaphoric links, it is also reinforced by a more general effect of the lexicon used in the text. On the one hand, different sentences in a cohesive text will tend to be about the same things. On the other hand, there will be patterns of word usage favouring the recurrence of previously used words, synonyms and other semantically related words. This aspect of cohesion is called *lexical cohesion* (Halliday and Hasan, 1976).

A somewhat related phenomenon in the context of translation is terminological consistency, which means that the same word will tend to be translated in the same way when it recurs in the text. Given a cohesive input text, this will help preserve cohesion under translation. Under the slogan of *one translation per discourse*, coined after the *one sense per discourse* hypothesis from computational semantics (Gale et al., 1992), this assumption was tested by Carpuat (2009) in a corpus study with both human translations and machine translations. She finds the hypothesis confirmed in the human translations in a corpus of English–French newswire. Perhaps more surprisingly, the hypothesis also holds in machine translations of the same texts generated by a phrase-based SMT system, an observation she puts down to the low variability of the SMT phrase tables. Of course, this result does not say anything about whether or not the consistent translations of the SMT system are correct.

We conclude that consistency of lexical choice is a property that cuts both ways. It is clearly a desired property of translated texts in some sense, but it may also be indicative of poor translation quality due to impoverished SMT

models. In the following sections, we try to shed some more light on this phenomenon. As a working hypothesis, we assume that SMT word choice could be improved by exploiting the vocabulary used in the whole text to make phrase selection consistent in the sense of lexical cohesion. In our experiments, we test if this is effectively the case under some operational models of lexical cohesion.

5.1.1 Translation Consistency in Different MT Systems

For our experiments, we use the English–French test set of the 2010 Metrics-MATR evaluation of MT evaluation metrics (Callison-Burch et al., 2010). The test set contains source, reference and the output of 22 different MT systems for a corpus of newswire text. To generate automatic word alignments between the source on the one hand and the reference and all candidate translations on the other hand, we concatenate the texts with the News commentary training corpus included in the WMT 2010 SMT shared task training data. Then we run GIZA++ (Och and Ney, 2003) in both translation directions and symmetrise the alignments with the `grow-diag-final-and` heuristic (Koehn et al., 2003).

The translations are first scored with a simple word translation model based on lexical weights, where the probability of a text is defined as follows:

$$L(T,S) = \log \prod_{s \in S} \frac{1}{|T_s|} \sum_{t \in T_s} p(t|s) \tag{5.1}$$

where S and T are the source and target language texts, s and t are single words and T_s is the set of target words aligned to a given source word. Unaligned words are considered to be aligned to a special NULL word. The probabilities $p(t|s)$ are estimated as unsmoothed relative frequencies computed over the text that is being scored. This score has the property of rewarding a consistent translation in which the same words are always translated in the same way.

The results of this experiment are shown in Fig. 5.1, where the lexical consistency score described in the previous paragraph is plotted against the percentage of acceptable translations according to the human evaluation for the WMT 2010 shared task (Callison-Burch et al., 2010). At the shared task, acceptability was determined with a two-stage procedure. In the first stage, the evaluators were asked to postedit the MT output in groups of five consecutive sentences to create fluent target language output, but without seeing either the source language input or the reference translations. In the second stage, they were asked to judge whether or not the postedited text was fully fluent in the target language and equivalent in meaning to the input text.

Under the model of Eq. 5.1, the reference translation is a clear outlier, and it obtains a *lower* score than any of the machine translations. This result

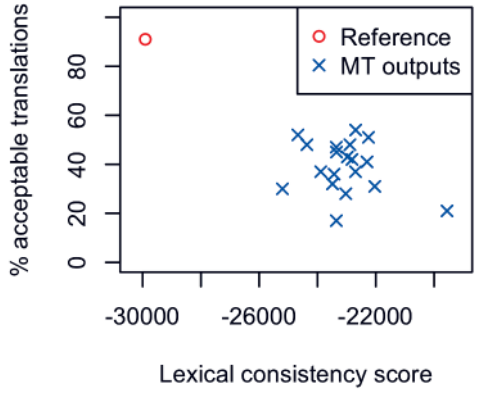

Figure 5.1. Lexical consistency vs. human MT evaluation for different MT systems

provides further evidence for the observation that SMT output uses fairly consistent vocabulary (Carpuat, 2009), but makes it appear improbable that a model of this kind can improve MT. Among the MT system outputs, there is no clear correlation between the lexical consistency score and the percentage of acceptable translations. A closer look at the individual systems and their system descriptions reveals that the differences in vocabulary consistency are due to other factors than just output quality; in particular, the size of the training corpus used for translation model and language model training has a large impact on the scores, smaller corpus size being correlated with higher consistency and lower translation quality. The presence of this nuisance variable makes it difficult to compare lexical consistency across different MT systems and confirms that excessive consistency may sometimes indicate poor translation quality, but it does not, of course, say anything about the usefulness of a lexical consistency or cohesion model when the training corpus size is kept fixed.

The conclusion that can be drawn from these experiments is that a model focusing on translation consistency alone is unlikely to improve SMT quality. Successful modelling of text cohesion will almost certainly require some source of semantic information.

5.1.2 Word-Space Models for Lexical Cohesion

In computational discourse modelling, word-space models generated by Latent Semantic Analysis (LSA) have been used to model the vocabulary consistency characteristic of lexical cohesion (Foltz et al., 1998; Beigman Klebanov et al., 2008; Gupta et al., 2008). By defining a lexical cohesion model on the basis of a word-space model, the cohesion model can be semantically anchored in a manner that is independent of the text to be scored. We hope that this kind of model will be able to distinguish between true lexical cohe-

sion and the delusive kind of consistency induced by the lack of variability in the SMT phrase tables.

As a preliminary experiment, we test a simple word-space cluster cohesion measure on the data set described in the previous section. We build a 300-dimensional word space model on French Wikipedia data using the LSA implementation found in the S-Space software package (Jurgens and Stevens, 2010). For each document in the translations produced by all MT systems as well as the human reference translator, the word vectors w_i of all words are looked up and averaged to determine the mean vector $\hat{w}$ for each document. Then, the score of an individual document is defined as the sum of squared distances between the individual word vectors and the document mean vector:

$$D = \sum_i |w_i - \hat{w}|^2 \tag{5.2}$$

The score of a test set is defined as the sum of the scores of all its documents. Note that in this experiment, unlike the previous one, a low score indicates high cohesion.

The results of this experiment are shown in Fig. 5.2. Unlike the simple consistency measure of the preceding section, according to which the reference translation seems to be less consistent than the MT outputs, the LSA-based measure judges that the reference, while being less of an outlier, is actually more cohesive than most of the machine-translated texts. Unfortunately, the diversity of the MT systems tested makes it difficult to draw more interesting conclusions. In particular, the only MT output with a lower score than the reference translation, indicating greater cohesion, comes from a shared task submission for which no system description paper was published, so it is unclear what properties of the system may have contributed to this result. The two submissions immediately following the reference translation in the score ranking (Federmann et al., 2010; Zeman, 2010) supply evidence that training corpus size has an effect on this score as well: These two systems use only a relatively small subset of the training data provided for the shared task. The system with the highest sum of squared distances is peculiar in a different way: Rather than using the training data provided by the shared task organisers, it is trained on a large corpus of training data extracted from translation memories of European Union translators (Jellinghaus et al., 2010).

5.1.3 A Semantic Document Language Model

We now present a model for lexical cohesion implemented in our document-level decoding framework. Our model rewards the use of semantically related words in the translation output by the decoder, where semantic distance is measured with a word space model based on Latent Semantic Analysis (LSA). LSA has been applied with some success to semantic language modelling

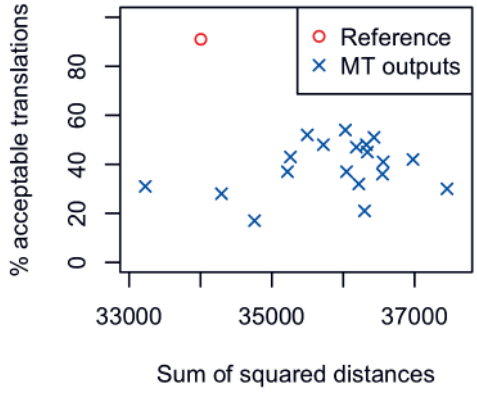

Figure 5.2. LSA cluster cohesion vs. human MT evaluation for different MT systems

in previous research (Coccaro and Jurafsky, 1998; Bellegarda, 2000; Wandmacher and Antoine, 2007). In SMT, it has mostly been used for domain adaptation (Kim and Khudanpur, 2004; Tam et al., 2007), or to measure sentence similarities (Banchs and Costa-jussà, 2011).

The model we use is inspired by Bellegarda (2000). It is a Markov model, similar to a standard n-gram model, and assigns to each content word a score given a history of n preceding content words, where $n = 30$ below. Content words are defined as tokens consisting exclusively of alphabetic characters not included in a stop word list originally developed for information retrieval (Savoy, 1999).[1] Scoring relies on a 30-dimensional LSA word vector space trained with the S-Space software (Jurgens and Stevens, 2010) on data from the Europarl and News commentary corpora of the 2010 WMT shared task. The score is defined based on the cosine similarity between the word vector of the predicted word and the mean word vector of the words in the history. Following Bellegarda (2000), we convert the similarity measure into a probability by looking at the empirical distribution of similarities between word vectors in the training set. The probability of a given similarity can then be estimated as the proportion of training examples having a lower similarity score than the target value.

The model is structurally different from a regular n-gram model in that word vector n-grams are defined over content words occurring in the word vector model only and can cross sentence boundaries. Stop words and tokens containing non-alphabetic characters, which together amount to around 60 % of the tokens, are scored by a different mechanism based on their relative frequency (undiscounted unigram probability) in the training corpus. In sum,

[1]The stop word list was retrieved from `http://members.unine.ch/jacques.savoy/clef/frenchST.txt` (12 October 2011).

Table 5.1. *Experimental results with a cross-sentence semantic language model*

	newstest2009		*newstest2010*		*newstest2011*	
	BLEU	NIST	BLEU	NIST	BLEU	NIST
DP search only	0.2256	6.513	0.2727	7.034	0.2494	7.170
DP + hill climbing	0.2260	6.518	0.2733	7.046	0.2497	7.169
with semantic LM	0.2271	6.549	0.2753	7.087	0.2490	7.199

the score produced by the semantic document LM has the following form:

$$h(w|h) = \begin{cases} p_{\text{unigr}}(w) & \text{if } w \text{ is a stop word, else} \\ \alpha p_{\cos}(w|h) & \text{if } w \text{ is a known word, else} \\ \epsilon & \text{if } w \text{ is an unknown word,} \end{cases} \tag{5.3}$$

where α is the proportion of content words in the training corpus and ϵ is a small fixed probability. It is integrated into the English–French SMT system described in Section 4.6 as an extra feature function for the Docent decoder. Its weight is selected by grid search over a number of values, comparing translation performance for the *newstest2009* test set.

In these experiments, we use DP beam search to initialise the state of our local search decoder. Three results are presented (Table 5.1): The first table row shows the baseline performance using DP beam search with standard sentence-local features only. The scores in the second row result from running the hill climbing decoder with DP initialisation, but without adding any models. A marginal increase in BLEU scores for all three test sets demonstrates that the hill climbing decoder manages to correct some of the search errors made by the DP search. The last row contains the scores obtained by adding in the semantic language model. Scores are presented for three publicly available test sets from recent WMT machine translation shared tasks, of which one (*newstest2009*) was used to monitor progress during development and select the final model.

Adding the semantic language model results in a small increase in NIST scores (Doddington, 2002) for all three test sets as well as a small BLEU score gain (Papineni et al., 2002) for two out of three corpora. We note that the NIST score proves to react more sensitively to improvements due to the semantic LM in all our experiments. This is reasonable because the model specifically targets content words, which benefit from the information weighting done by the NIST score. While the results we present do not constitute compelling evidence in favour of our semantic LM in its current form, they do suggest that this model could be improved to realise higher gains from cross-sentence semantic information and demonstrate how the document-level decoder enables experimentation with models that would be much more difficult to integrate in DP beam search.

5.2 Translating for Special Target Groups: Improving Readability

The experiments of the first half of this chapter were geared towards lexical cohesion, a phenomenon present in all connected text and universally relevant in translation. Discourse-level modelling can also be used to create output texts with certain specific properties that are desirable in a particular translation task. As an example, we consider models that improve the readability of the target text by exerting an influence on the vocabulary preferred by the SMT system and pushing it towards words and constructions that are potentially easier to understand.[2] This form of simplifying translation can be useful for special populations such as less proficient language users or dyslectic readers or simply for non-experts who want to grasp the main content in a domain-specific text, e. g., from the legal or medical domain, written in a foreign language.

Readability and text simplification have been widely studied in the field of computational linguistics, and several metrics and approaches have been proposed in the literature. Common readability metrics make use of global text properties such as type/token ratios, lexical consistency and the proportion of long versus short words. Our goal is to incorporate these features in an SMT system in order to combine text simplification and MT in a single system.

Chall (1958) identifies four main factors with strong effects on the readability of a text. Mühlenbock and Johansson Kokkinakis (2009) propose four corresponding quantitative indicators to measure them. *Vocabulary load* is the difficulty of the vocabulary; the corresponding measure is the number of words exceeding a certain length. *Sentence structure* is the syntactic complexity of the text and is measured by determining the average sentence length. *Idea density* represents the conceptual difficulty of the text, with lexical variation as a quantitative measure. Finally, *human interest* indicates the degree of abstractness of the text, and it is measured as the proportion of proper nouns. The proposed metrics are all fairly crude approximations of the motivating text qualities, but they have the advantage of being easy to measure and not requiring deep syntactic or semantic analysis.

5.2.1 Readability Metrics

The starting point for the work of Mühlenbock and Johansson Kokkinakis (2009) is a well-known readability metric for Swedish called LIX (*Läsbarhetsindex*; Björnsson, 1968). In the terminology of Chall (1958), the LIX metric covers the vocabulary load and the sentence structure dimensions. It is

[2]The results presented in this section are primarily the work of Sara Stymne (Stymne et al., 2013c), who carried out the experiments and composed an earlier version of the text on which this section is based.

computed as a linear combination of the average sentence length and the proportion of tokens longer than 6 characters, as in the following equation, where $C(x)$ is the count of x:

$$\text{LIX} = \frac{C(\text{tokens})}{C(\text{sentences})} + 100 \cdot \frac{C(\text{tokens} > 6 \text{ chars})}{C(\text{tokens})} \tag{5.4}$$

Average sentence length (ASL) is also useful as a standalone measure for sentence structure complexity:

$$\text{ASL} = \frac{C(\text{tokens})}{C(\text{sentences})} \tag{5.5}$$

Since complicated concepts are frequently expressed with long compound words in Swedish, Mühlenbock and Johansson Kokkinakis (2009) suggest measuring the percentage of extralong words with 14 characters or more as an additional indicator of high vocabulary load:

$$\text{XLW} = \frac{C(\text{tokens} \geq 14 \text{ chars})}{C(\text{tokens})} \tag{5.6}$$

Idea density could be measured with the type-token ratio:

$$\text{TTR} = \frac{C(\text{tokens})}{C(\text{types})} \tag{5.7}$$

In order to improve comparability across texts of different length, Hultman and Westman (1977, 56) propose a related, but different measure of lexical variation. They relate the vocabulary size or type count V of a text to its token count N as

$$V = N^{2-N^k} \tag{5.8}$$

for some text-specific constant k and define their lexical variation metric OVIX (*ordvariationsindex*) as the reciprocal of k. Solving for OVIX, we obtain:

$$\text{OVIX} = \frac{\log N}{\log\left(2 - \frac{\log V}{\log N}\right)} = \frac{\log C(\text{tokens})}{\log\left(2 - \frac{\log C(\text{types})}{\log C(\text{tokens})}\right)} \tag{5.9}$$

Another indicator of idea density is the nominal ratio NR:

$$\text{NR} = \frac{C(\text{nouns}) + C(\text{prepositions}) + C(\text{participles})}{C(\text{pronouns}) + C(\text{adverbs}) + C(\text{verbs})} \tag{5.10}$$

Typical news texts can be expected to have an NR around 1. NR correlates positively with formality and negatively with readability (Mühlenbock and Johansson Kokkinakis, 2009).

The proportion of proper nouns PN is used as an indicator of human interest:

$$\text{PN} = \frac{C(\text{proper names})}{C(\text{tokens})} \tag{5.11}$$

Finally, Stymne et al. (2013c) suggest that consistent translation may contribute to readability and propose using a measure of translation consistency based on an association metric called Q-score (Deléger et al., 2006). Q-score measures the association strength of an aligned pair of items and can be used either at the word level or at the level of SMT phrases. It is computed as the token frequency of the aligned pair *st* divided by the sum of the total number of pair types the source *s* and the target *t* individually occur in:

$$Q = \frac{C(st)}{N(s\diamond) + N(\diamond t)} \tag{5.12}$$

Here, $C(x)$ is the token frequency of x as above and $N(x)$ is the number of types matching a certain pattern, and the symbol $\diamond$ represents a wildcard character. Intuitively, the Q-score rewards common phrase or word pairs with consistent translations whereas it penalises less frequent pairs whose source and target elements also participate in many other pairs.

5.2.2 Experiments

For our experiments, we have implemented a subset of the readability features discussed in the previous section as feature functions for the Docent decoder described in Chapter 4. Some of the metrics can be evaluated at the sentence level, whereas others are meaningful only at the document level. The following features are implemented:

Sentence level	
SL	Sentence length in words
nLW	Number of long words (> 6 characters)
nXLW	Number of extralong words (≥ 14 characters)
Document level	
TTR	Type-token ratio (Eq. 5.7)
OVIX	Word variation index (Eq. 5.9)
Qw	Q-score, word level (Eq. 5.12)
Qp	Q-score, phrase level (Eq. 5.12)

We evaluate our models on parliamentary texts from the Europarl corpus (Koehn, 2005). This corpus contains both complex sentences and a great deal of domain-specific terminology. Our system is trained on 1,488,322 sentences of English–Swedish data. For evaluation, we extract 20 documents with a total of 690 sentences from a held-out part of Europarl. A document is defined as a complete contiguous sequence of utterances of one speaker. We exclude documents that are shorter than 20 sentences or longer than 79 sentences.

Moses (Koehn et al., 2007) is used for training the translation model and SRILM (Stolcke, 2002) for training the language model. We initialise our experiments with a Moses model that uses the standard features of a sentence-level phrase-based SMT system: a 5-gram language model, five translation

Table 5.2. *Systems with single readability features*

Feature	Weight	BLEU↑	NIST↑	LIX↓	ASL↓	OVIX↓	XLW↓	NR↓	PN↑
Reference	–	–	–	50.47	24.65	57.73	3.08	1.055	0.013
Baseline	–	0.243	6.12	51.17	25.01	56.88	2.63	1.062	0.015
OVIX	low	0.243	6.11	51.00	25.09	54.65	2.60	1.069	0.015
	medium	0.228	5.83	49.33	25.45	44.43	2.53	1.063	0.015
	high	0.144	4.41	46.59	29.09	31.65	1.82	0.941	0.013
TTR	low	0.243	6.12	51.04	25.11	55.25	2.60	1.070	0.015
	medium	0.225	5.75	49.86	26.19	45.31	2.44	1.080	0.014
	high	0.150	4.48	48.30	30.54	32.95	1.77	0.975	0.012
Qw	low	0.242	6.10	51.16	25.07	57.16	2.62	1.064	0.015
	medium	0.231	5.90	51.28	25.32	58.90	2.62	1.074	0.015
	high	0.165	4.93	50.92	26.14	60.61	2.63	1.101	0.016
Qp	low	0.243	6.12	51.16	24.99	56.94	2.65	1.061	0.015
	medium	0.229	5.99	49.79	24.14	54.75	2.62	1.060	0.015
	high	0.097	3.90	41.45	21.99	39.22	2.39	1.129	0.015
nLW	low	0.244	6.14	50.96	24.98	56.73	2.63	1.065	0.015
	medium	0.225	5.96	46.72	24.21	55.39	2.72	1.080	0.018
	high	0.106	4.11	30.27	22.18	45.41	1.78	0.899	0.023
nXLW	low	0.241	6.10	51.03	24.96	56.69	1.85	1.060	0.015
	medium	0.225	5.85	50.92	25.09	56.56	0.19	1.070	0.016
	high	0.224	5.84	50.97	25.12	56.55	0.19	1.068	0.016
SL	low	0.242	6.21	51.07	24.22	57.79	2.71	1.058	0.016
	medium	0.211	5.94	50.77	21.61	60.93	3.15	1.040	0.018
	high	0.150	4.38	50.77	18.46	65.37	3.72	1.072	0.021

↑ higher score is better ↓ lower score is better

model features, a distance-based reordering penalty and a word counter. The weights of these features are optimised using minimum error-rate training (Och, 2003). We reuse the same weights in Docent. The weights of the document-level features are not optimised automatically, not least because we have no tuning set with reference translations optimised for readability. Instead, we test three different settings with a low, a medium and a high weight relative to the other components of the weight vector for each readability feature.

Owing to the lack of other resources, we perform automatic evaluation against the standard reference translation contained in the Europarl corpus. This translation is in no way simplified or optimised for readability. We report figures for two standard MT evaluation scores, BLEU (Papineni et al., 2002) and NIST (Doddington, 2002), as well as the readability metrics discussed in the previous section.

Table 5.2 shows the results when we activate one readability feature at a time using low, medium, and high weights for each feature. The baseline and reference are quite similar with respect to readability with some interesting

Table 5.3. *Systems with combinations of readability features (medium weights)*

	BLEU↑	NIST↑	LIX↓	ASL↓	OVIX↓	XLW↓	NR↓	PN↑
Baseline	0.243	6.12	51.17	25.01	56.88	2.63	1.062	0.015
LIX (nLW+SL)	0.214	5.96	46.09	23.02	56.27	2.90	1.061	0.018
OVIX+SL	0.229	5.94	48.86	24.34	44.53	2.63	1.046	0.015
Qp+OVIX+nLW+SL	0.225	5.93	47.77	24.08	43.77	2.65	1.045	0.016
All features	0.235	6.04	49.29	24.34	47.80	1.98	1.046	0.015

differences, e. g., in the proportion of extra long words. As expected, giving a high weight to a readability feature usually results in a sharp decrease in MT quality with respect to the unsimplified reference translations, but it also greatly affects the corresponding readability features. In some cases, turning on the readability features results in extreme scores clearly indicative of overfitting. As an example, using the nLW feature with a high weight decreases the LIX score by more than 20 points.

Using low or medium weights, by contrast, can give reasonable MT scores as well as some improvements on several readability metrics. Unsurprisingly, the features corresponding directly to a metric, like the nLW feature for the LIX metric and the OVIX feature for the OVIX and TTR metrics, affect that metric strongly. Several features also have an effect on other readability metrics. For instance, the OVIX and TTR features improve several metrics, but cause an increase in sentence length, which is undesired. The effect of phrase-level Q-value is very different from that of word-level Q-value. On the phrase level, it improves most metrics, while its effect on the readability metrics is small when used on the word level.

In Table 5.3, we show results for some combinations of features, using medium weights. As expected, the effect on the readability metrics is more balanced in these cases. For the system with all features there are improvements on all readability metrics, except for PN, which is on a par with the baseline. The other systems that use some global feature also have a positive effect on most readability metrics, while the LIX system that uses only local features has little effect on OVIX and a negative effect on extra long words. For all these systems, the decrease in MT quality is modest. This shows that the decoder with its document-level features manages to simplify translations with respect to different aspects corresponding to vocabulary load, idea density, and sentence structure, while maintaining reasonable translation quality.

We also performed a small human evaluation of 100 random non-identical sentences from the baseline and the system using all readability features.[3] For each sentence we rank the output on adequacy, how well the content is translated, and readability, how easy to read the translations are. The results are shown in Table 5.4. The baseline produces a higher number of adequate

[3] 177 out of 690 sentences were identical.

Table 5.4. *Human preference with respect to adequacy and readability*

	Preferred system		
	Baseline	Equal	Readability (All)
Adequacy	51	33	16
Readability	33	29	38

translations than the system with readability features, but in many cases, adequacy is equal. For readability, there is a small advantage for the system with readability features, which is consistent with the improvement on readability metrics.

Overall, the output is often very similar with only a few words differing. In some of the cases where the baseline is judged as having better adequacy, the cause is a single changed word, which may be more common or shorter, but has the wrong form or part of speech, so it does not fit into the context. In other cases, some non-essential information is removed from the sentence, which while making the translation less adequate, is actually what we want to achieve. In some cases, however, the words removed from the translations do contain essential information.

In Table 5.5 (p. 86), we show some sample translations in order to exemplify the types of operations our current system is able to perform. One type of successful simplification is to remove words that are not crucial for the meaning of the text. Many of the systems with readability constraints simplify the phrase *the honourable Members*, either by removing the adjective and giving only *ledamöterna* 'the members', or even by using the pronoun *ni* 'you'. Another good simplification is the rendering of *in such a way that*, which is translated quite literally in the baseline, as *så att* 'so that' by several of the systems. There are also instances, however, where the changes lead to a loss of information. Examples are *handlingsplan* 'action plan', which is reduced to *plan* 'plan' by the nLW system, and *2003*, which is missing in the output of the OVIX and Qp systems. Often, different translations are chosen for a word or phrase. Sometimes this leads to a simplification, as in the nLW system, which uses the everyday expression *bli klar* 'finish' instead of the more formal *avsluta* 'finish'. In other cases, the translation options are of a relatively similar degree of difficulty, such as *vissa/en del/några* 'some', all of which are valid translations. In some cases the system with readability features prefers a translation with a different part of speech, as for *uppmärksamhet* 'attention', which is translated with the adjective *uppmärksam* 'attentive' by several systems. This leads to syntactic problems later on in the translation. In general, as can be expected of SMT, there are some problems with fluency in all translations, but they tend to get worse in the systems with high-weight readability features.

5.3 Conclusion

In sum, our experiments with the cross-sentence semantic models as well as with the readability models suggest that it is possible to control some output properties related to vocabulary choice with document-level features. Both types of models are in need of improvement.

The semantic language model introduced in Section 5.1.3 is a demonstration of a model that would be difficult, if not impossible, to implement with a sentence-level SMT decoder. It models n-gram-like sequences of content words that can span several sentences and introduces undirected dependencies between words that could not easily be processed with the two-pass or the sentence-to-sentence information propagation method. Of the techniques discussed in Chapter 3, only n-best rescoring could handle this kind of model, but it would be limited to a greatly impoverished representation of the search space.

The readability features we have explored show that our document-level MT system is capable of enforcing specific vocabulary properties in the output texts. In their current form, they have a negative effect on adequacy, even though they do improve the automatic readability scores. Overfitting to the scores is an aspect that must be considered in future development. It may also be necessary to reconsider the scores, most of which are just crude approximations of the relevant linguistic phenomena that are not necessarily correlated with perceived readability when systematically optimised against.

Another problem of the readability experiments we performed is the lack of relevant, simplified reference translations. This is a reflection of the fact that joint translation and text simplification strains the limits of the equivalence perspective on translation adopted by SMT. By applying readability models, the input text is retargeted to another audience, so the intentionality of the translation act can no longer be ignored. Evaluating against standard Europarl translations makes it difficult to assess the quality of simplified translations since any deviation from the reference will be penalised by the automatic evaluation scores, even if it is correct and has the desired effect of improving readability. Had we tried to tune the feature weights automatically, the same problem would have occured there, so the weights obtained with MERT using a regular, unsimplified test set are almost certainly suboptimal for use with the readability models. A corpus of simplified texts might also make it possible to adapt language models to simplified output, which in turn might improve fluency. Nevertheless, our experiments demonstrate the effectiveness of the decoding procedure introduced in Chapter 4.

Table 5.5. *Examples of translation output from systems with readability features*

Source	As the honourable Members know – some speakers have mentioned it – the European Council at Lisbon paid particular attention to promoting our efforts to implement risk capital in such a way that the action plan will be finished in 2003.
Baseline	Som de ärade ledamöterna vet – vissa talare har nämnt det – som Europeiska rådet i Lissabon ägnat särskild uppmärksamhet åt att främja våra ansträngningar att genomföra riskkapital på ett sådant sätt att handlingsplanen kommer att vara avslutat år 2003.
All (medium)	Som ledamöterna vet – vissa talare har nämnt det – som Europeiska rådet i Lissabon särskilt uppmärksam på att främja våra insatser för att genomföra riskkapital så att handlingsplanen kommer att vara avslutat 2003.
LIX (medium)	Som ledamöterna vet – vissa talare har nämnt det – Europeiska rådet i Lissabon lagt särskild vikt vid att främja våra ansträngningar att genomföra riskkapital så att handlingsplanen kommer att vara avslutat år 2003.
OVIX+SL (medium)	Som ni vet – vissa talare har nämnt det – som Europeiska rådet i Lissabon särskilt uppmärksam på att främja våra ansträngningar att genomföra riskkapital så att handlingsplanen kommer att avslutas under 2003.
OVIX (high)	Som ledamöter – en del talare har nämnt det – som Europeiska rådet i Lissabon särskilt uppmärksam på att stödja våra insatser för att genomföra av riskkapital, på så sätt att handlingsplanen kommer att vara avslutat i.
Qp (high)	Som de ärade ledamöterna vet, som några talare har nämnt det rådet i Lissabon, ägnat särskild uppmärksamhet åt att vi för att genomföra riskerna i det att handlingsplanen kommer att avslutas med.
nLW (high)	Som ni vet – vissa har sagt det – EU:s möte i Lissabon lagt särskild vikt vid vår för att genomföra risk i så att den plan att bli klar under 2003.
SL (high)	Som ledamöterna vet vissa talare har nämnt – Europeiska rådet i Lissabon särskilt uppmärksammat främja våra ansträngningar att genomföra riskkapital så att handlingsplanen avslutas 2003.

Part II:
Pronominal Anaphora in Translation

6. Challenges for Anaphora Translation

In this chapter, we introduce the problem of translating pronominal anaphora, which will be the main topic of the entire second part of this thesis. Pronominal anaphora is a specific discourse phenomenon that is ubiquitous in natural language text and poses surprisingly hard problems to SMT systems. In many languages, morphological features of anaphoric pronouns such as grammatical gender and number must agree with the corresponding features of their antecedents. Generating the right forms of the pronouns requires target-side dependencies because agreement depends on features in the linguistic system of the target language that do not necessarily map to properties of the input text. However, even though it seems obvious that it must be possible to improve SMT by considering anaphoric pronouns, both our own research and that of others have shown that it is far more difficult to obtain gains in translation quality than it might seem at first glance. We start by taking a closer look at what pronominal anaphora actually is and by establishing that it is in fact a problem for SMT. Then we discuss some of the difficulties that arise in recent work on pronouns in SMT.

6.1 Pronouns and Anaphora Resolution

Anaphora is "a relation between two linguistic elements, in which the interpretation of one (called an anaphor) is in some way determined by the interpretation of the other (called an antecedent)" (Huang, 2004). *Pronominal anaphora* specifically refers to the case in which the anaphor is a pronoun that should be interpreted as coreferring with something already mentioned. In the following example, the pronoun *them* in the second sentence has the same referent as and agrees morphologically with the noun phrase *the Catholics* in the first:

(6.1) *The Catholics* described the situation as "safe" and "protecting." This made *them* "relaxed and peaceful." (*newstest2009*)[1]

Prototypically, anaphoric pronouns refer to entities introduced into the discourse in the form of noun phrases. They can also refer to events or to parts of the discourse itself, or to phenomena not explicitly mentioned, but somehow implied by the discourse. Such cases are sometimes subsumed under the

[1]Examples marked *news-test2008* and *newstest2009* are taken from the 2008 and 2009 test sets of the WMT shared tasks, respectively (Callison-Burch et al., 2009).

label *event anaphora*. If a referring pronoun precedes its antecedent instead of following it, it is called *cataphoric* instead of anaphoric. Furthermore, some uses of pronouns do not refer to a particular antecedent at all. For instance, the *expletive* or *pleonastic* pronoun *it* in *it is raining* has a purely syntactic function and is not anaphoric.

Anaphora resolution, the problem of identifying the antecedent of an anaphoric linguistic element, is a long-standing research problem in computational linguistics. Much research has been devoted to noun phrase coreference resolution, which is "the task of determining which NPs in a text or dialogue refer to the same real-world entity" (Ng, 2010). For the purposes of this thesis, the special case of pronominal anaphora resolution is most relevant. General-purpose automatic coreference resolution systems usually try to resolve both pronominal and non-pronominal noun phrase coreference, whereas event anaphora tends to be somewhat neglected (Pradhan et al., 2011).

In many systems, automatic coreference resolution proceeds in two stages. First, the system analyses the text to be annotated with the help of NLP tools such as taggers or syntactic parsers and finds the noun phrases eligible for inclusion in a coreference relation, called *mentions* or *markables*. Then, it performs inference over the markables found to determine which of them refer to the same extra-linguistic *entities*.

There are different ways to approach the coreference resolution task. Ng (2010) distinguishes between *mention-pair systems* and *entity-mention systems*. The former try to decide, for each pair of mentions in the text, whether or not they refer to the same entity. The latter construct an abstract representation of all the entities in the text and decide, for each mention, whether or not it refers to a given entity.

Like MT, coreference resolution can be approached both with handwritten rules (e. g., Lee et al., 2011) or with machine learning methods. Systems whose core component is based on machine learning are often mention-pair systems using an extension of a basic set of 12 features originally proposed by Soon et al. (2001).

In many of the experiments contained in this thesis, we use the coreference system BART (Versley et al., 2008). BART is easily extensible and very modular, which makes it an excellent platform for our experimental work. Our version of BART is based on an official version released in 2010. Since then, the development of the coreference resolution system has continued, and it is likely that many features of our version do not correspond exactly to more recent releases of BART. Therefore, results involving coreference resolution that we present in this thesis should not be construed as reflecting the performance of current versions of BART.

Our version of BART has a mention-pair decoder with a set of features based mostly on the elementary feature set of Soon et al. (2001) and later work by Uryupina (2006). It has a mention detection pipeline that uses the

Morpha morphological analyser (Minnen et al., 2001), the Berkeley parser (Petrov et al., 2006; Petrov and Klein, 2007) and the Stanford named entity recogniser (Finkel et al., 2005). In the actual prediction component, the sentence containing the anaphoric pronoun and a limited number of sentences immediately preceding it are searched for markables that can serve as potential antecedents for the anaphor. Among these markables, the most probable candidate is selected with a maximum entropy ranker and returned.

6.2 Translating Pronominal Anaphora

When translating a discourse containing pronouns into another language, an MT system must decide how to render the input pronouns adequately in the target language. The choices that must be made for pronouns are potentially more difficult than when translating content words. To begin with, it is not even clear that every pronoun in the input should be translated into a corresponding pronoun in the translation. Mitkov and Barbu (2003) compare how the French translations of three technical texts written in English use pronouns compared to the originals. In their sample, the French translations contain almost 40 % more pronouns (390 instead of 281). For 241 pronouns, there is a $1 : 1$ correspondence between the languages, but 40 English pronouns and a staggering 159 French pronouns have either no direct correspondence or a corresponding full NP in the other language. Generalising these figures is problematic because the sample is small and it covers a very specific text type and only a single language pair and translation direction; furthermore, it is not known if the translations were created by the same or by different translators. In any case, the study clearly demonstrates that cross-lingual differences in pronoun use are by no means a marginal phenomenon.

For content words, MT systems usually assume that each item in the source text should be mapped into an equivalent item in the target language, possibly as an element of a multi-word phrase or idiom. Since suppression of content words would very likely entail a loss of information in the translation, this is a reasonable assumption to make for the literal translation style typical of MT, even though a human translator might sometimes opt for a less literal rendering of the input as a result of functional or pragmatic considerations. The use of pronouns, in contrast, is much more dependent on the linguistic structure and conventions of the target language, and it is by no means evident that an anaphoric pronoun should always be translated with an anaphoric pronoun even if fairly literal translation is sought for. For instance, when translating into languages like Italian or Spanish which do not require overt subject pronouns, English subject pronouns must be left out systematically to create a natural-sounding target text.

This is only one part of the problem, however. Even in the typical case, when an input pronoun is translated into a corresponding target language

pronoun, complications arise because many languages require agreement between the pronoun and its antecedent. The agreement relation must be enforced in the target language by considering the relevant features such as gender and number of the translation of the antecedent. Source language information found in the input is not sufficient alone to choose the correct pronoun. This is demonstrated by the following (contrived) example:

(6.2) a. The *funeral* of the Queen Mother will take place on Friday. *It* will be broadcast live.

b. Les *funérailles* de la reine-mère auront lieu vendredi. *Elles* seront retransmises en direct.

Here, the English antecedent, *the funeral of the Queen Mother*, requires a singular form for the anaphoric pronoun *it*. The French translation of the antecedent, *les funérailles de la reine-mère*, is feminine plural, so the corresponding anaphoric pronoun, *elles*, must be a feminine plural form too. Additionally, the French verbs are marked for plural in both sentences although the English verbs are singular forms. Consider, however, that the translator could have chosen to translate the word *funeral* with the perfectly correct French word *enterrement* 'burial' instead:

(6.3) L'*enterrement* de la reine-mère aura lieu vendredi. *Il* sera retransmis en direct.

Now, the antecedent NP is rendered as a masculine singular and correspondingly requires a masculine singular anaphoric pronoun and singular verb forms.

Importantly, there is nothing in the English source text to predict the gender of either the antecedent or the pronoun. English words do not have grammatical gender, but even if they did, it would not necessarily be predictive of gender in another language. Number marking will often be consistent across languages because it is more tightly knit to circumstances in the real world. Nonetheless, examples (6.2) and (6.3) show that discrepancies are possible for this feature as well. The only reliable predictor of the morphological features of a translated anaphoric pronoun is the translation of the antecedent, which, as the example illustrates, is to some extent at the discretion of the translator, or the MT system.

Anaphora is a very common phenomenon found in almost all kinds of texts. The anaphoric link can be local to the sentence, or it can cross sentence boundaries. In the first case, pronoun agreement may be dealt with correctly by the local dependencies of the SMT language model, but this becomes increasingly unlikely as the distance between the referring pronoun and its antecedent increases. The second, non-local case is not handled by standard SMT models at all. It is worth pointing out that a pronoun may well be translated correctly even without the benefit of a specific anaphora model because the SMT system easily learns an unconditional distribution

over the pronouns in the training sets. In example (6.1), the plural pronoun *them* would probably be rendered as *sie* by a naïve English–German SMT system, which is very likely to be a good choice. When translating into a language with gender-marked plural pronouns, however, selecting the right pronoun is more difficult.

6.3 A Study of Pronoun Translations in MT Output

To show that pronominal anaphora is indeed a problem for SMT, we study the performance of one of our SMT systems on personal pronouns. The sample examined in our case study is drawn from the German–English corpus used as a test set for the MT shared task at the EACL 2009 Workshop on Machine Translation (Callison-Burch et al., 2009). The test set is composed of 111 newswire documents from various sources in German and English translations. In the selected subset of 13 documents (219 sentences) we have identified all cases of pronominal anaphora that could be resolved in the text. One of the documents does not contain any such cases. For each anaphoric pronoun in the German source text, we manually check whether or not it was translated into English in an appropriate way by the phrase-based SMT system we submitted to the WMT 2010 shared task (Hardmeier et al., 2010). The system uses 6-gram language models, allowing it to consider a relatively large local context in translation, but it does not contain any specific components to process sentence-wide or cross-sentence context.

In this sample, the MT system finds a suitable translation for anaphoric pronouns in about 61 % of the cases (Table 6.1). How well it performs is strongly dependent on the type of pronoun: While it produces adequate output for around 90 % of the demonstrative pronouns (*dieser*, *dieses*, etc.) and about 3 out of 4 masculine or neuter singular pronouns or plural pronouns, only a third of the feminine pronouns are translated correctly. For pronouns of polite address and reflexive pronouns, the system largely fails.

The reasons for these discrepancies can most likely be found in the differences of the pronominal systems of the source and the target languages. The English system of pronouns distinguishes between human (*he, she*) and non-human (*it*) referents in the singular. A gender distinction is made only for humans. The German nominal system has three grammatical genders, which do not correspond directly to biological sex and apply also to inanimate objects. They are distinguished in the singular forms of the pronouns.

Moreover, some German pronouns are highly ambiguous. Thus, the pronoun *sie* can be the form of the feminine singular, of the plural of any gender or, when spelt *Sie* with an uppercase initial letter, of the polite form of address, which is usually translated into an English second person *you*. The reflexive pronoun *sich* is used for all genders and both numbers in the third person; it frequently has no direct equivalent in the English sentence. In these ambigu-

ous cases, the language model will try to disambiguate based on parts of the context that were seen during training. If the local context is truly ambiguous, the results of the disambiguation will be essentially random. Generally, the system will prefer the forms that were observed most frequently at training time. For instance, given the pronoun distribution in typical corpora of newswire text and political speeches, it will tend to translate *sie* as a plural pronoun even when it is a feminine singular in reality. As a result of these factors, pronoun translation accuracy varies greatly from document to document according to the number and types of pronouns that occur.

Even though translation mistakes due to wrong pronoun choice generally do not affect important content words, they can make the MT output hard to understand, as in the following example from document 3 of our sample:

(6.4) a. *Input:* Der Strafgerichtshof in Truro erfuhr, dass *er seine* Stieftochter Stephanie Randle regelmässig fesselte, als *sie* zwischen fünf und sieben Jahre als [recte: alt] war.

b. *Reference translation:* Truro Crown Court heard *he* regularly tied up *his* step-daughter Stephanie Randle, when *she* was aged between five and seven.

c. *MT output:* The Criminal Court in Truro was told *it* was *his* Stieftochter Stephanie Randle tied as *they* regularly between five and seven years. (*newstest2009*)

There are several things wrong with this MT output, and bad pronoun choice is clearly one of them, with the pronoun *er* referring to a male person translated as *it* and the pronoun *sie* referring to a female person translated as *they*.

To sum up, there is evidence that current phrase-based SMT cannot handle pronoun choice adequately. Although our case study is limited to a single language pair and a single text genre, considering the models used in SMT, there is no reason to suppose that the situation should be very different in other cases. Stronger differences in pronoun systems and text with longer, more complex sentences are likely to exacerbate the difficulties, whereas the problem will be easier to solve when the languages are close and the sentences are simple and match the training corpus closely.

6.4 Challenges for Pronoun Translation

The results of the case study in the previous section indicate that better handling of pronominal anaphora may lead to observable improvements in translation quality. However, the attempts at explicit pronoun modelling for SMT reported in the literature (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2011; Hardmeier et al., 2013b) suggest that the problem is harder than it seems. Pronoun translation is a complex task, and solving it correctly requires a number of steps, including identification of anaphoric

Table 6.1. *Correct translations and total number of German anaphoric pronouns in a subset of the WMT 2009 test set.*

Document		masc. sg.	fem. sg.	neuter sg.	plural	polite address	reflexive	demonstrative	pron. + prep.	total	
1	Aktualne.cz	1/ 1	–/ 1	–/ 1	–/ 1	–/ –	–/ 2	1/ 1	–/ 2	2/ 9	22 %
2	Spiegel	–/ –	–/ –	–/ –	–/ –	–/ –	–/ –	–/ –	–/ –	–/ –	–
3	BBC	5/ 8	6/23	1/ 2	–/ –	1/ 4	–/ 4	2/ 2	–/ –	15/ 43	35 %
4	BBC	9/11	1/ 2	2/ 2	–/ –	–/ –	–/ –	–/ –	1/ 1	13/ 16	81 %
5	Times	1/ 3	2/ 2	–/ –	7/10	–/ –	–/ –	1/ 1	–/ –	11/ 16	69 %
6	ABC.es	7/13	–/ 1	1/ 1	3/ 3	–/ –	1/ 1	2/ 3	–/ –	14/ 22	64 %
7	El Mundo	4/ 5	2/ 3	8/ 8	–/ –	–/ –	–/ –	4/ 4	–/ –	18/ 20	90 %
8	Les Echos	2/ 3	–/ –	–/ –	–/ –	–/ –	–/ –	–/ –	–/ –	2/ 3	67 %
9	Le Devoir	16/19	2/ 8	4/ 4	2/ 2	–/ –	1/ 2	3/ 3	–/ –	28/ 38	74 %
10	hvg.hu	2/ 2	–/ –	1/ 4	4/ 4	–/ –	1/ 2	2/ 2	–/ –	10/ 14	71 %
11	nemzet.hu	–/ –	1/ 6	–/ –	–/ –	–/ –	–/ –	–/ –	–/ –	1/ 6	17 %
12	Adnkronos	–/ –	2/ 2	–/ –	1/ 2	–/ –	–/ –	2/ 3	–/ –	5/ 7	71 %
13	Corriere	2/ 3	–/ 1	–/ –	–/ –	–/ –	–/ –	1/ 1	–/ –	3/ 5	60 %
		49/68	16/49	17/22	17/22	1/ 4	3/11	18/20	1/ 3	122/199	61 %
		72 %	33 %	77 %	77 %	25 %	27 %	90 %	33 %	61 %	

pronouns, correct translation of the parts of the discourse containing the antecedents, recognition of the anaphoric link to the right antecedent, extraction of relevant features from the antecedent, generation of the correct pronoun and its embedding in a correct translation of its context. Each of these steps is in itself non-trivial, and there is a substantial risk that noise introduced by errors in each part of the task accumulates and eradicates all useful information in the chain.

Guillou (2012) discusses a number of reasons for the disappointing performance of SMT systems with anaphora handling. In particular, she identifies four main sources of error:

1. Identification of anaphoric vs. non-anaphoric pronouns,
2. Anaphora resolution,
3. Identification of the head of the antecedent noun phrases, from which gender and number features are extracted,
4. Word and phrase alignment between source and target text.

While we largely agree with Guillou's analysis of these problems, we believe that the list should be extended. We have identified six principal factors that present risks to pronoun-aware SMT systems and may help to explain the failure of existing research to find solutions:

1. Baseline SMT performance,
2. Anaphora resolution performance,
3. Performance of other external components,
4. Inadequate evaluation,
5. Error propagation, and
6. Model deficiencies.

The sources of error listed by Guillou (2012) can be subsumed under these headings. In the following sections, we examine these challenges in more detail, beginning with risks external to the pronoun translation approaches proper and continuing with deficiencies inherent in the methods that were tested in the literature. From this discussion, we derive the insights that shaped the key features of our recent work presented in the later chapters of this thesis (Chapters 8 and 9; Hardmeier et al., 2013b).

6.4.1 Baseline SMT Performance

Models for anaphoric pronouns target a very specific linguistic phenomenon by manipulating a small number of words in the output text. This can only be successful if the translation as a whole is reasonably good; no pronoun translation model will achieve significant improvements if what the underlying SMT system outputs without its help is mostly gibberish. It is well known that some language pairs are much more difficult for SMT than others, for

instance because of word order differences or complex target language morphology. In other cases, out-of-vocabulary words in the input text may make the translation unreliable. When this happens, there is not much that a pronoun model can do to improve the translation because it is too specifically focused on a single phenomenon.

In our English–German system (Hardmeier and Federico, 2010), we experienced insufficient baseline performance as a major problem. Similarly, Guillou (2011) remarks that "[o]ne of the major difficulties that [human evaluators] encountered during the evaluation was in connection with evaluating the translation of pronouns in sentences which exhibit poor syntactic structure." This suggests that, at least in some cases, the translations output by her English–Czech MT system were so poor as to render pronoun-specific evaluation essentially meaningless.

By contrast, the output of state-of-the-art English–French SMT systems is to a large extent intelligible if not perfect. It sometimes happens that the SMT system garbles the syntax of a sentence, such as in the following examples, where the words of the input sentence are reordered in a manner that completely distorts the meaning of the sentences:

(6.5) a. *Input:* We don't have stewardesses, we've been against it from the very beginning.

b. *MT output:* Nous n'avons pas, nous avons été hôtesses contre elle dès le début. (*newstest2009*)

(6.6) a. *Input:* And this time, Hurston's old neighbors saw her as a savior.

b. *MT output:* Et cette fois, l'ancienne Hurston voisins a vu son comme un sauveur. (*newstest2009*)

In comparison to other language pairs, these cases are fairly rare, however, and it is reasonable to assume that this was the case also for the anaphora-sensitive English–French systems described in the literature (Le Nagard and Koehn, 2010; Hardmeier et al., 2011). Generally, there is little researchers interested in anaphora can do about this problem except working on an easier language pair while waiting for the progress of general SMT research.

6.4.2 Anaphora Resolution Performance

Any MT system that attempts to model pronominal anaphora explicitly must identify anaphoric links in the input in some way, be it by running a separate anaphora resolution component (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010), by performing anaphora resolution jointly together with pronoun prediction (Hardmeier et al., 2013b) or by relying on manual gold-standard annotations (Guillou, 2011). When many anaphoric links are resolved incorrectly, a model may degrade performance on average rather than improve it. To see why, consider that an SMT system with no explicit ana-

phora handling component will not emit pronouns randomly; rather, the system is likely to have a preference for the pronouns that are most frequent in the training corpus. If the test set is homogeneous with the training data, this may very well be the correct choice in many cases.

As an example, the SMT system used in pronoun translation corpus study described above (Section 6.3; Hardmeier et al., 2010) has a strong preference for translating the ambiguous German pronoun *sie* as *they* or *them* rather than *she* or *her*. In consequence, pronoun translation errors are very frequent in documents whose main character is female, whereas many other documents are hardly affected. Clearly, this is a problem not only from a technical, but also from a gender-political point of view (Gendered Innovations, 2014). Overall, anaphora resolution is a difficult task in itself, and inadequate performance of the coreference resolver has been advanced as an explanation for disappointing experimental results in at least one study (Le Nagard and Koehn, 2010).

Pronouns are notoriously difficult for anaphora resolution systems to resolve correctly when they do not refer to a noun phrase. On the one hand, this applies to expletive pronouns such as *it* in *it is raining*, which are not used anaphorically at all. Detecting expletives automatically is a hard problem. Le Nagard and Koehn (2010) implement a rule-based system for this task (Paice and Husk, 1987), which performs surprisingly well for them at a precision and recall of 83 %; however, the same system has been shown to perform considerably worse on different corpus data (Evans, 2001). One of the best systems currently available, achieving high accuracy on a variety of test sets, is the one by Bergsma and Yarowsky (2011).

Low recall for expletive classification means that a substantial part of the expletive pronouns in a text will be incorrectly linked to an antecedent. As an example, consider the following two sentences, where the version of the BART coreference resolution system used by Hardmeier et al. (2011) incorrectly links the non-referring pronoun *It* in the second sentence to the word *it* in the first and creates a coreference chain *price – it – It*:

(6.7) Napi's basket suggested that this latter was a near impossibility, since we found that the price was up by just a shade over 10 percent on last year's quite high base *price*, even where *it* was most expensive. *It* does appear, though, that flour suppliers are in a stronger position than egg producers, for they have managed to force their drastic price increases onto the multinationals. (*news-test2008*)

On the other hand, pronouns may refer to an event expressed by a verb phrase rather than to a noun phrase, as in the following example:

(6.8) He made a scandal out of *it* when the Prefecture ordered the dissolution of the municipal council. (*newstest2009*)

This type of coreference is handled less consistently by current coreference resolution systems (Pradhan et al., 2011), so pronouns with event anaphora will often be resolved incorrectly as referring to a noun phrase. At the same time, both expletives and event anaphora may be relatively easy for a naïve SMT system to get right, since they are generally rendered with a small set of common pronouns such as *it* in English or *il*, *ça*, *cela* in French. In such cases, incorrect anaphora resolution greatly increases the risk of mistranslation.

6.4.3 Performance of Other External Components

Recognising and resolving pronominal anaphora in a document and transferring it into another language requires analysis at a relatively high level of linguistic abstraction. Depending on the architecture of a specific system, a variety of external components may be used to perform certain steps of this analysis. In addition to the potentially quite complex preprocessing pipelines of their coreference resolution systems, existing systems (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010) rely on external resources to identify morphological features of potential antecedents and to align the words of the source language to those of the target language. While these are well-researched NLP tasks and good tools exist, their accuracy is not perfect, and all errors will add to the level of noise present in the total system.

Tools for morphological analysis are language-specific and will not be available for all languages in the same quality. Even for a language like French that may well have one of the best collections of NLP tools after English, it turns out to be surprisingly difficult to obtain a reliable morphological analyser that works well on all text types. A number of systems have been developed, but not all of them are publicly available and perform adequately on the MT test corpora. Both Le Nagard and Koehn (2010) and Hardmeier and Federico (2010) use the Lefff full-form lexicon (Sagot et al., 2006). This is an excellent resource with wide, but obviously not perfect, coverage, and as a pure lexicon resource it contains multiple analyses of some ambiguous word forms. To words not listed in the dictionary, Hardmeier and Federico (2010) apply a small number of rule-based heuristics that improve coverage somewhat. Still, the quantity of words with no or an incorrect analysis is not negligible, and these words may provoke translation errors.

Cross-lingual word alignment is an essential step in the SMT training process. The development of statistical alignment methods stands at the very beginning of this research field (Brown et al., 1990, 1993). The success of SMT relies strongly on the accuracy of these methods, but also on the tolerance of subsequent training steps to the errors they make. When training translation models for phrase-based SMT, word-to-word alignments are used as the basis for an elaborate heuristic phrase extraction procedure (Och and Ney, 2004) that extracts all phrase pairs consistent with a word alignment

according to certain criteria. This method copes very effectively with word alignment errors by reducing the influence of individual alignment links. Frequently, phrase pairs will be identified correctly even though some word in them is not aligned, or incorrectly aligned. A pronoun translation model cannot have the same kind of tolerance because it must consider the alignment links of individual words. What is more, pronouns, which it is particularly interested in, may be particularly prone to having erroneous alignment links. Since they are very common and are not translated strictly literally in many cases, they have fairly high translation probabilities to all kinds of words in the word alignment models. As a result, linking them to other common nearby words often increases the alignment score even if the correspondence is not motivated linguistically. In the worst case, they may get aligned to a totally unrelated pronoun in the other language, so that the pronoun translation model enforces an incorrect translation for that pronoun.

6.4.4 Inadequate Evaluation

It is widely recognised that automatic evaluation of pronoun translation is difficult and existing methods are unreliable (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2011). Popular MT evaluation metrics such as BLEU (Papineni et al., 2002) score the MT output by comparing it to one or more reference translations. This approach is fraught with problems. Since it is completely unspecific and assigns the same weight to any overlap with the reference, it is not particularly sensitive to the type of improvements targeted by a pronoun translation component, which affect only a few words in a text.

Hardmeier and Federico (2010) address this shortcoming by using a precision/recall-based measure counting the overlap of pronoun translations in the MT output and a reference translation (see Chapter 7 for details). Whilst increasing the sensitivity to pronoun changes, this measure retains another serious drawback of a reference-based pronoun evaluation in that it judges correctness by comparing the translation of a pronoun in the MT output with the translation found in a reference translation and assumes that they should be the same. However, this assumption is flawed: It does not necessarily hold if the MT system selects a different translation for the antecedent of the pronoun. If this is the case, the only meaningful way to check the correctness of a pronoun is by finding out whether it agrees with the antecedent selected by the system, even if the translation of the antecedent may be incorrect.

As Guillou (2011) remarks, the usefulness of an evaluation method that checks pronouns against a reference translation also depends on the number of inflectional forms for pronouns in the target language. If pronouns are inflected for a large number of features in a given language, the probability of matching a pronoun exactly with a noisy system is very low even if many

of its features are generated correctly, and it becomes difficult to measure progress before perfection is achieved.

More relevant conclusions about the quality of pronoun translation could be drawn by examining how the MT output renders the coreference chains found in the input and checking the pronouns referring to the same entity for consistency. The main difficulty here is that this makes the evaluation dependent on coreference annotations for the source language, leading to unreliable evaluation results when there are errors in the annotation. This evaluation strategy was adopted by Guillou (2011) and worked well for her since she had gold-standard coreference annotations for her test set. In the absence of gold-standard annotations, reliable evaluation of pronoun translations seems difficult or impossible. Coreference-annotated parallel corpora like the Prague Czech–English Dependency Treebank (Hajič et al., 2006) and the recently developed ParCor corpus containing data for English–French and English–German (Guillou et al., 2014) are essential resources for sound evaluation of pronoun translations.

6.4.5 Error Propagation

In the definition cited above (Section 6.1), anaphora is defined as "a relation between two linguistic elements, in which the interpretation of one (called an anaphor) is in some way determined by the interpretation of the other (called an antecedent)" (Huang, 2004). This definition focuses on the linguistic realisation of the anaphor and the antecedent, and it views anaphora as a pairwise relation between exactly two linguistic elements. This focus is shared by other definitions of the terms *anaphora* (Bussmann, 1996) and *anaphor* (Trask, 1993). In the case of nominal coreference and pronominal anaphora, it could be argued, however, that the immediate relation holds not between two linguistic elements, but between a linguistic element and an entity in the real world, or the representation of an entity in the reader's or listener's mind, which was presumably evoked by one or more linguistic elements in the preceding discourse. It could also be argued that the anaphoric relation holds between the anaphor and the set of all linguistic elements referring to the same entities.

The formal representation of anaphoric links in a computational system must commit to one of these views. In coreference resolution, it is common to encode anaphoric links as coreference classes, defined as the set of all mentions in a document referring to the same entity. The extratextual, non-linguistic nature of the entities is emphasised by the definition of NP coreference resolution as "the task of determining which NPs in a text or dialogue refer to the same real-world entity" (Ng, 2010). This is consistent with the last of the definitions mentioned above. In many practical implementations, however, the anaphoric link is represented primarily as a pairwise

relation between two noun phrases in the text, a view more compatible with the encyclopaedic definitions referred to first. These pairwise links are then usually converted into coreference classes for evaluation.

In an anaphora model for SMT, it is often easier to deal with pairwise anaphoric links than with entire coreference classes, especially if one of the sentence-based decoding procedures described in Chapter 3 is applied. To some extent, this is also justified because the morphological agreement relation, with which anaphora models for SMT are mostly concerned, holds between the anaphoric pronoun and the most recent, or possibly most salient, mention in the text, not between the pronoun and an abstract concept. In the existing literature, mention-pair representations of anaphoric links are practically universal (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2011). Conditioning the translation decision for an anaphoric pronoun on the translation of a single antecedent NP creates a risk of error propagation. This is particularly relevant if a coreference chain consists of a sequence of pronouns. If the SMT system, triggered by some other factor such as the *n*-gram model, mistranslates one of the pronouns in the chain, this error can easily be propagated to all later elements of the chain. This problem could be addressed by processing the coreference links so that links pointing to an antecedent that is a pronoun are transitively extended until a full NP is reached, but even in this case, the presence of a single incorrect link may lead to false resolution and, consequently, false pronoun choice.

6.4.6 Model Deficiencies

Le Nagard and Koehn (2010) claim that "[their] method works in principle," if it wasn't for the poor performance of the coreference resolution system, and Hardmeier and Federico (2010) report minor improvements for the pronoun *it* in a pronoun-specific automatic evaluation with their method. However, later work suggests that both methods are in need of refinement before they can deliver consistently useful results by demonstrating that performance remains unconvincing even when using gold-standard coreference annotations (Guillou, 2011) and that the small improvements that have been realised do not carry over to another language pair (Hardmeier et al., 2011).

An interesting observation made by both Guillou (2011) and Hardmeier et al. (2011) is that SMT systems with explicit pronoun handling tend to generate more pronouns than required. The reason for this need not be the same for both systems. In particular, in the English–Czech system, one difference between the languages is that Czech, unlike English, allows subject pronouns to be left out when the subject can be inferred from the context. The observed overgeneration effect may result from a reduced tendency of the second-pass system with its more focused pronoun translation distributions to drop pro-

nouns, word removal being an event not explicity accounted for in the standard phrase-based SMT model.

In the experiments by Hardmeier et al. (2011), anaphoric links are modelled by a bigram language model predicting pronouns given gender and number of the antecedent. The vocabulary of the predicted words is restricted to pronominal forms. Other words are treated as "out of vocabulary" by the model and penalised harshly. This leads to a strong preference for translating every single pronoun as a pronoun, even when this is not an adequate translation, e. g., when the coreference system mistakenly resolved a non-referential pronoun by linking it to an antecedent.

In sum, the existing pronoun models for SMT are clearly less than perfect, and pronoun overgeneration is a problem that has been observed repeatedly with different models. To improve the models, the reasons for this behaviour should be examined more closely. It may be necessary to design an explicit model for dropping pronouns or translating them with non-pronouns. As pointed out earlier, research on anaphora resolution has had a tendency towards focusing on the prototypical case of anaphora with a nominal antecedent, and non-referential pronouns and event anaphora pose harder challenges to current systems. The same preference for prototypical problem instances can be observed in research on SMT pronoun models; in SMT, however, the less frequent, non-prototypical cases may in fact be easier to handle for a naïve system since, at least for target languages like French or German, agreement patterns are much less complex than for nominal antecedents. Consequently, there is a substantial risk of degrading performance by adding a pronoun model that mishandles these very categories.

6.5 Conclusion

In the previous section, we gave an overview of the main challenges that an SMT system with an explicit pronoun model is faced with. The analysis we presented is a result of our earlier work on pronoun translation, some of which we present in the following chapter. The insights gained from this work have influenced our more recent work on pronouns, which will be the topic of the remainder of this thesis. Let us therefore recapitulate the challenges discussed above and consider the design decisions we have made to cope with them.

The first factor we mentioned is baseline performance, which means the performance of all components of the SMT system except the ones we are interested in. What we can do here is select our baseline system so as to maximise the effect of the model we want to test. For pronoun translation, it seems important to choose a language pair with very good SMT performance as it is almost impossible to improve on an underperforming MT system with a pronoun model. At the same time, it is important that there should be an

interesting difference in pronoun systems between the source and the target language.

For us, baseline performance was the main reason to give up language pairs such as German–English and German–French, which we studied in earlier work. Even though these language pairs are very interesting from the point of view of pronoun translation, the word order differences between German on one side and English and French on the other, as well as the relatively complex morphology of German, make it difficult to train good phrase-based SMT systems. Instead, we concentrate our efforts on the language pair English–French. This is a combination of two major European languages with plenty of resources. Both languages have very simple noun morphology, and their word order is very similar. At the same time, there is an interesting difference between the French third person pronouns, which follow a two-gender system that conflates biological and grammatical gender for both animate and inanimate entities, and the English pronouns, which are marked for animacy but do not have gender features on inanimate pronouns.

Many of the difficulties related to coreference resolution, morphological analysis, error propagation and pronoun modelling in general are addressed in our work on pronoun prediction described in Chapter 8. Our design decisions are guided by the modelling assumptions outlined in Section 1.3. One of the most important consequences of our early experiments is that we try to reduce our dependence on external tools and integrate as much of the task as possible into our own system. To the maximum extent possible, we avoid pipeline architectures in favour of tightly integrated components. Thus, the neural network classifier we present in Chapter 8 combines pronoun prediction with anaphora resolution in a single network. Tight coupling permits us to preserve the uncertainty of the individual steps; rather than resolving a pronoun to a single antecedent, we propagate a set of antecedent candidates with an associated probability distribution to the next step. Doing so should also reduce the risk of error propagation a little by minimising the effect of uncertain decisions, even if it does not solve its root cause and coreference chains are still modelled as sequences of pairwise links.

Uniting different parts of the task in one system allows us to train the entire system in one go for a single training criterion that matches the objective of pronoun prediction for which the classifier will finally be used, and it ensures that all parts of the system are trained on the same type of training data. The alternative would often be to train components such as anaphora resolution systems or morphological analysers on out-of-domain data because annotated training data for the target domain may not be available. It has been shown at least for word sense disambiguation that matching the training objectives and data sets of an SMT system and its ancillary components can be essential for success (Carpuat and Wu, 2007). We suggest that this may be a factor for pronoun translation too.

Evaluation is the problem we contribute least to in this work. In the next chapter, we briefly discuss a pronoun-specific evaluation metric that is based on precision and recall of pronoun translation, but it is still unsatisfactory and suffers from many of the same weaknesses as the existing, general evaluation measures. In Chapter 9, we present a method for annotating and evaluating pronoun translations in SMT output, which allows us to analyse the performance of our own anaphora model. Parallel coreference-annotated data for the English–French language pair has only been developed very recently (Guillou et al., 2014) and was unavailable for most of the work contained in this thesis. In our experiments in Chapter 9, we use this new resource as a source of reliable anaphora annotations for our model, but the development of better evaluation measures must be left to future work.

7. A Word Dependency Model for Anaphoric Pronouns

In this chapter, we describe some of our early results on pronominal anaphora translation. We present a simple document-level word dependency model for the Moses decoder and its application to pronominal anaphora for the language pair English–German (Hardmeier and Federico, 2010). It represents one of the earliest attempts to integrate knowledge about pronominal anaphora into the standard, sentence-level tools of phrase-based SMT. The initial publication of this work was one of the very first papers that addressed the problem of pronominal anaphora in SMT (together with Le Nagard and Koehn, 2010). We also introduce an evaluation metric that specifically measures the accuracy of pronoun translation and is more sensitive to the effects of our anaphora models on the MT output than standard automatic MT evaluation measures such as BLEU (Papineni et al., 2002).

To enable discourse-level information processing for our word dependency model in a sentence-level SMT framework, we apply the sentence-to-sentence information propagation approach described in Section 3.4. Anaphoric links are modelled as directed dependencies between word pairs consisting of a pronoun and its closest antecedent. Links are identified with the help of an external coreference resolution system. Our model assigns a probability to the translation of a pronoun given the translation of its antecedent. It handles both sentence-internal and cross-sentence anaphora.

7.1 Anaphoric Links as Word Dependencies

In general, the decision what translation to emit in the target language for a given source pronoun cannot be taken based on local information only. In many languages, pronouns show complex patterns of agreement, and selecting the correct word form requires dependencies on potentially remote words. German possessive pronouns, for instance, agree in gender and number with the possessor (determining the choice between *sein*, *ihr*, etc.) and in gender, number and case with the possessed object (with a paradigmatic choice between, e. g., *sein*, *seine*, *seines*, etc., if the possessor is masculine singular). While the possessed object occurs in the same noun phrase as the pronoun and agreement can, at least in simpler cases, be enforced by an *n*-gram language model, the possessor can occur anywhere in the text, even in

[The same **hospital**]$_1$ had had to contend with a similar infection early this year. [**It**]$_{2 \rightarrow 1}$ had discharged a patient admitted after a serious traffic accident. Shortly afterward, [**it**]$_{3 \rightarrow 2}$ had to re-admit the patient because of an MRSA infection, and [**doctors**]$_4$ have been unable to perform surgery that would be vital to full recovery because [**they**]$_{5 \rightarrow 4}$ have been unable to get rid of the staph.

```
The same hospital had had to contend with a similar infection early
this year .
It|*->neut_sg had discharged a patient admitted after a serious
traffic accident .
Shortly afterward , it|*->neut_sg had to re-admit the patient because
of an MRSA infection , and doctors|1-* have been unable to perform
surgery that would be vital to full recovery because they|*-1 have
been unable to get rid of the staph .
```

Figure 7.1. Coreference link annotation and decoder input

a different sentence. Since a given input word can be translated with different words in the target language and the pronoun must agree with the word that was actually chosen, correct pronoun choice depends on a translation decision taken earlier by the MT system. Our model extends the SMT decoder with the capacity to handle dependencies between the translations of words regardless of their distance in the input. The relevant word pairs are identified by an external anaphora resolver, and the objective of the model is to promote morphological agreement between anaphoric pronouns and their antecedents.

We use the open-source coreference resolution system BART (Versley et al., 2008) to link pronouns to their antecedents in the text. The coreference resolution system was trained on the *ACE02-npaper* corpus (Mitchell et al., 2003) and uses separate models for pronouns and non-pronouns in order to increase pronoun-resolution performance. For each resolvable pronoun, the system finds a link to an antecedent NP. Exactly one NP per pronoun is found, and it is the closest NP preceding the pronoun that the anaphora resolver considers as coreferent with the pronoun. Our word dependency model handles links between pairs of individual words, not syntactic phrases, so we identify the syntactic head of the antecedent NP with the Collins head finder (Collins, 1999) and represent the anaphoric relation as a link between the anaphoric pronoun and the syntactic head word of its antecedent NP. The output of the coreference resolver is illustrated in the upper part of Fig. 7.1. Markable NPs are enclosed in square brackets and their syntactic heads are highlighted in bold face. After identifying direct anaphoric links, the coreference resolution system proceeds to cluster mentions into coreference chains, but we do not use this information in our experiments.

We integrate coreference information into an SMT system based on the phrase-based Moses decoder (Koehn et al., 2007) in the form of a new model

which represents dependencies between pairs of target-language words produced by the MT system. The decoder driver encodes the links found by the coreference resolver in the input passed to the SMT decoder. Pronouns and their antecedents are marked as illustrated in the lower half of Fig. 7.1. Each token is annotated with a pair of elements. The first part numbers the antecedents to which there is a reference in the same sentence. The second part contains the number of the sentence-internal antecedent to which this word refers, or a representation of the relevant features of the word itself, if it occurred in a previous sentence. Each part can be empty, in which case it is filled with an asterisk.

To reduce vocabulary size and data sparseness, we map the antecedent words to a tag representing their gender and number. In the example, the word *hospital* in the first sentence, which is translated by the system into the neuter singular word *Krankenhaus* (not shown), gets mapped to the tag `neut_sg` in the input for sentence 2. Gender and number of German words were annotated using the RFTagger (Schmid and Laws, 2008). The representation of the pronouns, by contrast, is fully lexicalised.

7.2 The Word Dependency Model

The word dependency module is integrated as an additional feature function in a standard SMT model (Eq. 3.1). It keeps track of pairs of source words $(s_{\text{ant}}, s_{\text{pron}})$ participating as antecedent and anaphor in a coreference link. Usually, the antecedent s_{ant} will be processed first; however, it is also possible for the anaphor s_{pron} to be encountered first, either because of a cataphoric link in the source sentence or, more likely, because of word reordering during decoding. When the second element of an antecedent-anaphor pair is translated, the word dependency module adds a score of the following form:

$$p(T_{\text{pron}}|T_{\text{ant}}) = \max_{(t_{\text{pron}}, t_{\text{ant}}) \in T_{\text{pron}} \times T_{\text{ant}}} p(t_{\text{pron}}|t_{\text{ant}}), \tag{7.1}$$

where T_{pron} is the set of target words aligned to the source word s_{pron} and T_{ant} is the set of target words aligned to the source word s_{ant} in the decoder output. Word alignments between decoder input and decoder output are constructed based on the phrase-internal word alignments computed during SMT system training.

Coreference links across sentence boundaries are handled by the decoder driver module of Section 3.4. It reads the decoder output and extracts the required information about antecedents occurring in previous sentences, encoding it in the input of the sentence containing the reference as described above. In the cross-sentence case, the antecedent is not marked in the decoder input, but once it has been translated, its translation is silently extracted from the output, and the anaphor token is decorated directly with the

gender/number tag corresponding to the extracted word form. Cataphoric links across sentence boundaries are not handled by the model.

In the DP search algorithm of a standard phrase-based SMT decoder, two search paths can be recombined if one of them is provably superior to the other under every possible continuation of the search (see Section 3.2). Since our model introduces dependencies that can span large parts of the sentence, care must be taken not to recombine hypotheses that could be ranked differently after including the word dependency scores. We therefore extend the decoder search state to include, on the one hand, the set of antecedents already processed and, on the other hand, the set of anaphors encountered for which no antecedent has been seen yet. In either case, the translation chosen by the decoder is stored along with the item. Hypotheses can only be recombined if both of these sets match.

Training our word dependency model requires estimating the conditional probability distribution $p(t_{\text{pron}}|t_{\text{ant}})$ in Eq. 7.1. We do so by computing relative frequencies in a training corpus and applying standard language model smoothing methods. Training examples are extracted from a parallel corpus in a way similar to the application of the model: The source-language part of a word-aligned parallel corpus is annotated for coreference with the BART software, then the antecedent and anaphor words are projected into the target language using the word alignments and the corresponding pairs of target-language antecedent and anaphor words are used as training examples. Apart from removing the need for an anaphora resolution system for the target language, using the source language system for both the training and testing stage has the advantage of greater consistency, but training the model directly on coreference pairs extracted in the target language would be a plausible alternative.

Our model is trained on version 10 of the News commentary corpus from the training data for the WMT 2010 shared task. The estimated probabilities are smoothed using the Witten-Bell method (Witten and Bell, 1991). This smoothing method does not make prior assumptions about the distribution of n-grams in a text. It is therefore more suited for estimating the probabilities of events not drawn directly as n-grams from a text than the improved Kneser-Ney method (Chen and Goodman, 1998) we use for smoothing our other n-gram models.

7.3 Evaluating Pronoun Translation

Assessing the quality of pronoun translation in SMT output with standard MT evaluation methods is problematic for several reasons. All widely used automatic evaluation metrics for MT measure the similarity between a candidate translation and one or more reference translations. The quality of a candidate translation is assumed to correlate with its similarity to the refer-

ence translations. Regardless of how similarity is defined, this can be no more than an approximation because any source text generally admits of a large variety of translations into a given target language.

The most popular automatic MT evaluation metric is certainly the BLEU score (Papineni et al., 2002). It measures the similarity between a candidate translation and a set of reference translations by looking at *n*-grams, usually of length up to 4 words, and counting how large a proportion of the *n*-grams in the candidate translation are found in the references too. When computing this *n*-gram precision quantity, BLEU uses *clipped n-gram counts* for the candidate translation. Clipping the counts means that every *n*-gram in the candidate translation is counted at most as often as the same *n*-gram occurs in a single reference translation. It makes sure that the MT system cannot inflate its score artificially by generating a great number of very common words that are likely to occur in many references. This is the formal definition of the clipped counts, with $c_C(N)$ being the count of *n*-gram N in the candidate translation and $c_R(N)$ its count in any reference translation R:

$$c_{\text{clip}}(N) = \min\left(c_C(N), \max_R c_R(N)\right) \tag{7.2}$$

Precision is calculated by summing up the clipped counts of all *n*-grams in the candidate translation and dividing by the total number of *n*-grams in the candidate. This quantity is multiplied by a *brevity penalty* that ensures that the MT system cannot optimise precision by suppressing all words it is not confident about. Essentially, the brevity penalty replaces a measure of recall. It is used because it is not straightforward to define recall when there are multiple reference translations.

For the evaluation of pronoun translation, BLEU has several important drawbacks. One of them is its total lack of specificity. BLEU assigns the same weight to any type of token, content word, function word, pronoun, verb, conjunction and punctuation mark alike. We are specifically interested in pronouns, but the BLEU score conflates pronouns with all kinds of other words and gives us a figure that may have little to do with what we actually want to measure.

Another limitation of BLEU is that it does not check whether an *n*-gram in the candidate translation actually corresponds to the *n*-gram it is matched with in the reference translation. In the case of content words, this may work well enough. If both the candidate translation and a reference translation contain the same highly informative and relatively rare word, the chances that they correspond to each other are fairly good. For common function words such as pronouns, however, the assumption breaks down. The fact that two translations both contain the word *it*, or *and*, or a comma sign, says little about their resemblance, unless the sentences are very short.

Finally, there is an even more serious issue with a similarity score like BLEU that makes it unsuitable for evaluating pronoun translation. BLEU as-

sumes that any overlap of the candidate translation with the reference translation is a sign of good quality, whereas any difference indicates poor quality. However, an anaphoric pronoun is correct only if it agrees with its antecedent. If the candidate translation renders the antecedent with an expression that does not match the reference, then the pronoun may have to be different and the pronoun of the reference translation may in fact be incorrect. If, say, an antecedent that is masculine in the reference translation is rendered with a feminine NP in the candidate, a simple similarity score will behave inconsistently and assign a higher score to a translation referring to the feminine antecedent with a masculine pronoun than to one having the correct feminine pronoun because the latter will be penalised for two mismatches with the reference translation instead of one despite being more grammatical.

We now present a simple method to measure the accuracy of pronoun translations more directly. Compared to BLEU, our method addresses the first two of the issues mentioned above by focusing specifically on pronouns, ignoring other word classes, and by using word alignments to keep track of the role of pronouns in a sentence to avoid conflating unrelated items as BLEU does. Like BLEU, however, it matches the translations of pronouns against a reference translation and does not solve the last problem we discussed.

We use a test corpus with a single reference translation. We construct word alignments for the candidate translation and the reference translation by concatenating them with additional parallel training data, running the GIZA++ word aligner (Och and Ney, 2003) in both directions and symmetrising the alignments as is usually done for SMT system training. We also produce word alignments between the source text and the candidate translation by considering the phrase-internal word alignments stored in the phrase table. The basic idea of our metric is to count the number of pronouns translated correctly. Doing so would require a $1 : 1$ mapping from pronouns to their translations. However, word alignments can link a word to zero, one or more words, so we suggest using a measure based on precision and recall instead.

For every pronoun occurring in the source text, we obtain the set of aligned target words in the reference and the candidate translation, R and C, respectively. Inspired by the BLEU score, we define the clipped count of a particular candidate word w as the number of times it occurs in the candidate set, limited by the number of times it occurs in the reference set:

$$c_{\text{clip}}(w) = \min\left(c_C(w), c_R(w)\right) \tag{7.3}$$

We then consider the match count to be the sum of the clipped counts over all words in the candidate translation aligned to pronouns in the source text, which allows us to define precision and recall in the usual way:

$$\text{Precision} = \frac{\sum_{w \in C} c_{\text{clip}}(w)}{|C|} \qquad \text{Recall} = \frac{\sum_{w \in C} c_{\text{clip}}(w)}{|R|} \tag{7.4}$$

This measure can be applied either to obtain a comprehensive score for a particular system on a test set or to compute detailed scores per pronoun type to gain further insights into the workings of the model.

For testing the significance of recall differences, we use a paired t-test. Pairing is done at the level of the set R, the individual target words aligned to pronouns in the reference translation. This method is not applicable to precision, as the sets C cannot be paired among different candidate translations.

7.4 Experimental Results

The baseline system for our experiments was built for the English–German task of the ACL 2010 Workshop on Statistical Machine Translation. It is a phrase-based SMT system based on the Moses decoder with phrase tables trained on version 5 of the Europarl corpus and version 10 of News commentary corpus and a 6-gram language model trained on the monolingual News corpus provided by the workshop organisers. The language model is estimated with modified Kneser-Ney smoothing (Chen and Goodman, 1998) using the IRSTLM language modelling toolkit (Federico et al., 2008).

The feature weights are optimised by running MERT (Och, 2003) against the *news-test2008* development set for the baseline system. In order to minimise the influence of feature weight selection on the outcome of the experiments, we do not rerun MERT after adding the word dependency model. Instead, we reuse the baseline feature weights and conduct a grid search over a set of possible values for the weight of the word dependency model, selecting the setup that yields best pronoun translation F-score on *news-test2008*. The weight is set to 0.05 with the other 14 weights (7 distortion weights, 1 language model, 5 translation model weights and word penalty as in a baseline Moses setup) normalised to sum to 1.

English–German is a relatively difficult language pair for SMT because of pervasive differences in word order and very productive compounding processes in German. Our baseline system achieves a BLEU score of 0.1366 on the *newstest2009* test set. The best system submitted to WMT 2009 scores 0.148 on the same test set. Handling pronouns with a word dependency model has no significant effect on the BLEU scores, which vary between 0.136 and 0.137 in all our experiments.

The pronoun-specific evaluation (Table 7.1) suggests that the SMT system is very bad at translating pronouns in general. Most of the pronoun translations do not match the reference. For both test sets, adding the word dependency model results in a tiny improvement in precision and a small improvement in recall, which is however highly significant ($p < .0005$ in a one-tailed t-test for both test sets).

A closer look at the performance of the system on individual pronouns reveals that by far the largest part of the improvement stems from the pro-

Table 7.1. *Pronoun translation precision and recall*

	news-test2008	*newstest2009*
Baseline		
Precision	0.333	0.428
Recall	0.302	0.388
F1	0.317	0.407
Word-dependency model		
Precision	0.338	0.430
Recall	0.316	0.399
F1	0.326	0.414

noun *it*, which is translated significantly better by the enhanced system than by the baseline. Recall for this pronoun improves from 0.210 to 0.271 for the *news-test2008* corpus ($p < .0001$, two-tailed t-test) and from 0.218 to 0.251 for the *newstest2009* corpus ($p < .005$). The only other item with a significant improvement at a confidence level of 95 % is, surprisingly enough, the first-person pronoun *I* in the *newstest2009* corpus (from 0.604 to 0.624, $p < .05$). In the *news-test2008* corpus, the word dependency model has no effect whatever on the word *I*, so it seems likely that this improvement is accidental.

By contrast, the improvement we obtain for the pronoun *it*, albeit slight, is encouraging. While most other English pronouns such as *he*, *she*, *they*, etc. are fairly unambiguous when translated into German and the ambiguity the MT system is faced with will mostly concern case marking or the difficult question whether or not a pronoun is to be translated as a pronoun at all, translating *it* requires the system to determine the grammatical gender of the German antecedent in order to choose the right pronoun. Similar problems occur in the opposite translation direction and in other language pairs, e. g., when translating the highly ambiguous German pronoun *sie* into English, or when translating between two languages that have different systems of grammatical gender. However, when applying our pronoun translation model to the language pair English–French, we do not observe any improvement at all either in the BLEU score or in the pronoun-specific evaluation score (Hardmeier et al., 2011).

7.5 Conclusion

Together with the two-pass approach by Le Nagard and Koehn (2010), the word dependency model described in this chapter was one of the first attempts to model pronominal anaphora in statistical MT (Hardmeier and Federico, 2010). A key property shared by both of these early approaches is that they try to make maximum use of existing tools and technologies and com-

bine them for a new purpose while making as little changes to their inner workings as possible. Our word dependency model is a straightforward extension of a standard sentence-level phrase-based SMT decoder, and most of the document processing logic is implemented outside the decoder. For coreference resolution, we completely rely on an external tool. Even the word dependency model itself is trained with standard language modelling software. Delegating most of the work to various external tools has the advantage of relative simplicity and can be implemented with limited effort. Unfortunately, it turns out to be quite difficult to achieve translation quality gains in this way.

Without going into much detail, we note that our word dependency model and the SMT system it was tested with suffer from many of the issues discussed in Chapter 6. To begin with, the difficulty of creating a good baseline system for translating from English into German makes it hard to achieve strong results with a pronoun translation component. Even so, the fact that we obtained no better results when we applied the same system to English–French translation with a much stronger baseline (Hardmeier et al., 2011) proves that this is not the only issue. The performance of the external coreference system and the quality of the gender and number annotations were additional problems. While we did not conduct a formal evaluation of these components, it was easy to see that there was a substantial level of noise in these annotations.

The most serious shortcomings, however, can be found in the word dependency model itself. The score of this model is calculated as a simple conditional probability that formally corresponds to a bigram language model score and is computed with language modelling tools. The antecedent, the element the probability is conditioned on, is represented as a gender/number tag, whereas the anaphoric pronoun is represented as a lexical item. This setup is unsatisfactory for several reasons. The antecedent encoding contains very little information. Hard decisions are made both when resolving the anaphoric link and when annotating the antecedent with its gender/number tags. Both types of annotations are subject to noise and errors, but the word dependency model knows nothing about the confidence with which the decisions were made. The word dependency model itself, on the other hand, is probabilistic and trained on noisy data. Because of errors made during the preparation of the training data, there will be a considerable number of training examples with combinations of antecedent tags and pronouns that do not agree morphologically. As a result, a substantial part of the probability mass is spilt on incorrect combinations that are mere artefacts of the training process.

Furthermore, in many cases source language pronouns are not aligned to pronouns in the target language, so the model score will be calculated based on a word that is not a pronoun at all. If the target language word has not been seen aligned to an input pronoun during training, it will be treated as an unknown word by the LM library and penalised strongly, promoting overgen-

eration of pronouns. This is an effect we observe in the translations output by the system.

Finally, anaphoric links are represented as pairwise relations between an anaphoric pronoun and its antecedent. The coreference resolution system prefers to link the pronoun to its closest antecedent, even if the antecedent is itself a pronoun. Pronoun-pronoun links are susceptible to errors because there is little information in the two pronouns to guide the anaphora resolver. As a result, a single incorrect link may introduce an error into a chain of pronouns with the effect that all subsequent pronouns get translated incorrectly. A similar situation can occur even if all anaphoric links are resolved correctly because of the stochastic nature of the word dependency model. Since some probability estimates for non-agreeing tag/pronoun pairs may be inflated as described in the preceding paragraph, errors may be stochastically introduced in a pronoun chain and propagated onwards.

To sum up, the word dependency model presented in this chapter suffers from a number of problems. It was one of the earliest attempts to model anaphora translation in SMT, and it has been useful because we have gained a better understanding of the difficulties hidden in the seemingly innocuous task of pronoun translation by identifying and studying its deficiencies. These insights have been material to the development of the models presented in the remainder of this thesis.

8. Cross-Lingual Pronoun Prediction

In the previous chapter, we discussed a simple word dependency model to represent anaphoric links in phrase-based SMT and demonstrated that its effect on pronoun translation was minimal and insufficient from the point of view of translation quality. We now leave aside the generation of translations for a while. Instead, we focus on the automatic prediction of pronoun translations when the surrounding discourse and its translation are known and cast pronoun translation as a classification task. Initial experiments with a simple maximum entropy classifier quickly reveal that classification is made difficult by the uneven distribution of personal pronouns. It is easy to achieve moderately good overall performance just by frequently predicting the most frequent classes, but this comes at the cost of very low recall for less frequent items such as the French feminine plural pronoun *elles*. A classifier with such characteristics is unlikely to improve SMT quality because it exhibits the same bias as a baseline SMT system without any pronoun-specific components. We propose a neural network classifier that achieves more consistent precision and recall and manages to make reasonable predictions for all pronoun categories in many cases.

We then go on to extend our neural network architecture to include anaphoric links as latent variables. We demonstrate that our classifier, now with its own source language anaphora resolver, can be trained successfully with backpropagation. In this setup, we no longer use the machine learning component included in the external coreference resolution system (BART; Versley et al., 2008) to predict anaphoric links. Instead, we rely on the additional information contained in our parallel training corpus to draw inferences about anaphoric relations. Anaphora resolution is done by our neural network classifier and requires only some quantity of word-aligned parallel data for training, completely obviating the need for a coreference-annotated training set.

8.1 Task Setup

The overall setup of the pronoun prediction task is shown in Fig. 8.1. We are given an English discourse containing a pronoun along with its French translation and word alignments between the two languages, which in our case were computed automatically using IBM model 4 (Brown et al., 1993) as implemented by GIZA++ (Och and Ney, 2003) and word alignment symmetrisation with the `grow-diag-final-and` heuristic (Koehn et al., 2003). We

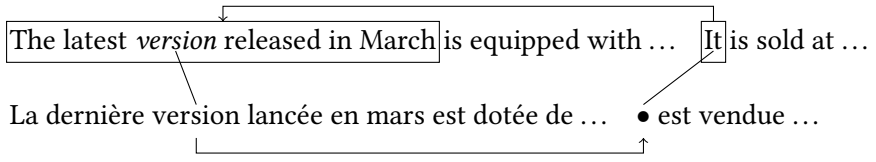

Figure 8.1. Task setup

focus on the four English third-person subject pronouns *he*, *she*, *it* and *they*. Note that the pronoun *it*, unlike the other pronouns, can also be an object pronoun, which adds a certain amount of noise to our data sets. The output of the classifier is a multinomial distribution over six classes:

- Four classes corresponding to the four pronouns *il*, *elle*, *ils* and *elles*. These are the masculine and feminine singular and plural forms of the third person subject pronoun, respectively.
- One class corresponding to the impersonal pronoun *ce* or *c'*, which occurs in some very frequent constructions such as *c'est* 'it is'. The elided form *c'* is used when the following word starts with a vowel. For the purpose of our classifier, we treat it as identical to the full form.
- A sixth class OTHER, which indicates that none of these pronouns was used.

In general, a pronoun may be aligned to multiple words. In this case, a training example is counted as a positive example for a class if the target word occurs among the words aligned to the pronoun, irrespective of the presence of other aligned tokens.

This task setup resembles the problem that an SMT system must solve to make informed choices when translating pronouns, but it avoids dealing with automatically generated target language text and uses human-made translations as target language context instead. This could make the task both easier and more difficult; easier, because the context can be relied on to be correctly translated, and more difficult, because human translators frequently create less literal translations than an SMT system would.

The features used in our classifier come from two different sources:

- *Anaphora context features* describe the source language pronoun and its immediate context consisting of three words to its left and three words to its right. They are encoded as vectors whose dimensionality is equal to the source vocabulary size with a single non-zero component indicating the word referred to (one-hot vectors).
- *Antecedent features* describe an antecedent candidate. Antecedent candidates are represented by the target language words aligned to the syntactic head of the source language markable noun phrase as identified by the Collins head finder (Collins, 1999).

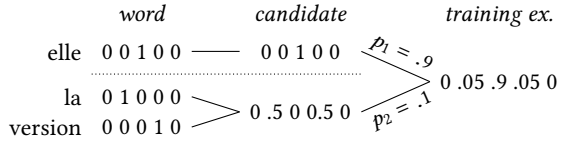

Figure 8.2. Antecedent feature aggregation

The encoding of the antecedent features is illustrated in Fig. 8.2 for a training example with two antecedent candidates translated to *elle* and *la version*, respectively. The target words are represented as one-hot vectors with the dimensionality of the target language vocabulary. These vectors are then averaged to yield a single vector per antecedent candidate. Finally, the vectors of all candidates for a given training example are weighted by the probabilities assigned to them by the anaphora resolver (p_1 and p_2) and summed to yield a single vector per training example.

The different handling of anaphora context features and antecedent features is due to the fact that we always consider a constant number of context words on the source side, whereas the number of antecedent word vectors to be considered depends on the number of antecedent candidates and on the number of target words aligned to the head word of each antecedent.

8.2 Data Sets and External Tools

We run experiments with two different test sets. The TED data set consists of around 2.6 million tokens of lecture subtitles released in the WIT^3 corpus (Cettolo et al., 2012). We extract 71,131 training examples from this corpus. The examples are randomly partitioned into a training set of 56,905 examples and a validation set and a test set of 7,113 examples each. For the maximum entropy classifiers described in the next section, another implementation of the extraction procedure is used, which differs in some edge cases. It yields 71,052 examples, randomly partitioned into a training set of 63,228 examples and a test set of 7,824 examples. The official WIT^3 development and test sets are not used in our classifier experiments because we want to reserve some held-out data for MT experiments.

The News commentary data set is version 6 of the parallel News commentary corpus released as a part of the WMT 2011 training data. It contains around 2.8 million tokens of news text and yields 31,090 data points, which are randomly split into 28.090 training examples and validation and test sets of 1,500 examples each. The extraction procedure for maximum entropy classifiers extracts 31,017 data points, randomly split into 27,900 training examples and 3,117 test instances.

Table 8.1. *Distribution of classes in the training data*

	TED		News commentary	
ce	6,901	16.3 %	1,312	6.4 %
elle	3,574	7.1 %	2,513	10.1 %
elles	1,581	3.0 %	995	3.9 %
il	8,645	17.1 %	5,865	26.5 %
ils	8,259	15.6 %	3,669	15.1 %
OTHER	42,171	40.9 %	16,736	38.0 %
	71,131	100.0 %	31,090	100.0 %

Table 8.2. *Percentages of French pronouns aligned to English pronouns*

	TED				News commentary			
	he	*she*	*it*	*they*	*he*	*she*	*it*	*they*
ce	1.0	1.1	15.3	1.6	1.0	0.6	6.3	1.6
elle	0.2	57.9	3.6	0.6	0.1	55.9	11.9	1.2
elles	–	0.1	0.2	8.4	–	0.4	0.5	10.5
il	54.6	0.4	9.7	2.1	55.1	1.4	18.8	2.9
ils	0.2	–	0.3	45.5	0.0	–	1.2	40.2
OTHER	44.0	40.5	70.9	41.8	43.8	41.7	61.4	43.6
	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

The distribution of the classes in the two training sets is shown in Table 8.1. One thing to note is the dominance of the OTHER class, which pools together such different phenomena as translations with other pronouns not in our list (e. g., *on* or *celui-ci*) and translations with full noun phrases instead of pronouns. Splitting this group into more meaningful subcategories is not straightforward, and it is even unclear if it would benefit performance because less frequent categories may be used in more varied ways while training data becomes ever sparser.

Table 8.2 shows how the examples in the two training sets are distributed among the different class labels. Although the two corpora belong to fairly different text genres, the distributions are similar. The most notable exceptions concern the translations of *it*. In the TED data, it is most frequently aligned to *ce*, indicating that the *c'est* 'it is' construction is very common in this corpus. The feminine *elle* is relatively infrequent. In the News corpus, translations with *il* are much more common at the expense of *ce*. This probably reflects a difference in modality and formality between the two corpora, the TED corpus being less formal and representing an oral genre. By contrast, the pronoun *elle* referring to feminine antecedents is more frequent as a translation of *it* in the News commentary corpus.

Table 8.3. *Majority class baseline results*

	TED (Accuracy: 0.622)			News commentary (Accuracy: 0.555)		
	P	R	F	P	R	F
ce	–	0.000	–	–	0.000	–
elle	0.579	0.536	0.557	0.559	0.111	0.185
elles	–	0.000	–	–	0.000	–
il	0.546	0.481	0.511	0.553	0.383	0.453
ils	0.455	0.985	0.622	–	0.000	–
OTHER	0.556	0.881	0.682	0.709	0.711	0.710

The feature setup of all our classifiers requires the detection of potential antecedents and the extraction of features pairing anaphoric pronouns with antecedent candidates. Some of our experiments also rely on an external anaphora resolution component. We use the open-source anaphora resolver BART, which we also used in the experiments of the previous chapter, to generate this information.

In all the experiments of this chapter, we use BART's markable detection and feature extraction machinery. In the experiments of the next two sections, we also use BART to predict anaphoric links for pronouns. The model used with BART is a maximum entropy ranker trained on the *ACE02-npaper* corpus (Mitchell et al., 2003). In order to obtain a probability distribution over antecedent candidates rather than one-best predictions or coreference sets, we have modified the ranking component with which BART resolves pronouns to normalise and output the scores assigned by the ranker to all candidates instead of picking the highest-scoring candidate. This is motivated by the observation that the correct antecedent is often assigned a relatively high score even if the single top-scoring candidate is incorrect. By preserving the uncertainty of the anaphora resolver's decision for the next steps in the pipeline, the effect of incorrect decisions should be mitigated. A drawback of this method, however, is that the BART model used was not trained in this condition, so the resulting probabilities may not be well calibrated.

8.3 Baseline Classifiers

The easiest way to create a reasonable baseline for our pronoun prediction task is to predict the majority class output for each source pronoun. This means that we always predict *il* for *he*, *elle* for *she* and OTHER for *it*. For *they*, both *ils* and OTHER are common in both corpora, but the optimal majority class prediction is *ils* for the TED corpus and OTHER for the News comment-

Table 8.4. *Maximum entropy classifier results*

	TED (Accuracy: 0.685)			News commentary (Accuracy: 0.576)		
	P	R	F	P	R	F
ce	0.593	0.728	0.654	0.508	0.294	0.373
elle	0.798	0.523	0.632	0.530	0.312	0.393
elles	0.812	0.164	0.273	0.538	0.062	0.111
il	0.764	0.550	0.639	0.600	0.666	0.631
ils	0.632	0.949	0.759	0.593	0.769	0.670
OTHER	0.724	0.692	0.708	0.564	0.609	0.586

aries. Table 8.3 shows the results for these predictions. In this and all the following tables, the label P corresponds to precision, R to recall and F to balanced F-score, the harmonic mean of precision and recall. Since the distributions are heavily skewed, the overall accuracy of this classifier is well over 50 % despite the number of output classes. The pronouns *ce* and *elles*, as well as *ils* in the News commentary corpus, are minority choices for all source pronouns, so they are never generated at all. In the TED corpus, there are comparatively more personal pronouns referring to humans, so *il* and *elle* are more frequently generated by *he* or *she*. This explains why the baseline scores for these pronouns are higher.

As a more sophisticated baseline, we train a maximum entropy (ME) classifier with the MegaM software package[1] using the features described in the previous section and the anaphoric links found by BART. The results are shown in Table 8.4. The F-scores are consistently over the majority class baseline for all pronouns and both corpora. As before, the overall accuracy is higher for the TED data than for the News commentary data. While precision is above 50 % in all categories and considerably higher in some, recall varies widely.

The pronoun *elles* is particularly interesting. This is the feminine plural of the third person subject pronoun, and it usually corresponds to the English pronoun *they*, which is not marked for gender. In French, *elles* is a marked choice which is only used if the antecedent is exclusively comprised of linguistic elements of feminine grammatical gender. The presence of a single item with masculine gender in the antecedent will trigger the use of the masculine plural pronoun *ils* instead. This distinction cannot be predicted from the English source pronoun or its context; making correct predictions requires knowledge about the antecedent of the pronoun. Moreover, *elles* is an infrequent pronoun. There are only 1,909 occurrences of this pronoun

[1] `http://www.umiacs.umd.edu/~hal/megam/` (20 June 2013).

in the TED training data, and 1,077 in the News commentary training set. Because of these special properties of the feminine plural class, we argue that the performance of a classifier on *elles* is a good indicator of how well it can represent relevant knowledge about pronominal anaphora as opposed to overfitting to source contexts or acting on prior assumptions about class frequencies.

In accordance with the general linguistic preference for *ils*, the classifier tends to predict *ils* much more often than *elles* when encountering an English plural pronoun. This is reflected in the fact that *elles* has much lower recall than *ils*. Clearly, the classifier achieves a good part of its accuracy by making majority choices without exploiting deeper knowledge about the antecedents of pronouns.

An additional experiment with a subset of 27,900 training examples from the TED data confirms that the difference between TED and News commentaries is not just an effect of training data size, but that TED data is genuinely easier to predict than News commentaries. In the reduced data TED condition, the classifier achieves an accuracy of 0.673. Precision and recall of all classifiers are much closer to the large-data TED condition than to the News commentary experiments, except for *elles*, where we obtain an F-score of 0.072 (P 0.818, R 0.038), indicating that small training data size is a serious problem for this low-frequency class.

8.4 Neural Network Classifier

In the previous section, we saw that a simple multiclass maximum entropy classifier, while making correct predictions for much of the data set, has a significant bias towards making majority class decisions, relying more on prior assumptions about the frequency distribution of the classes than on antecedent features when handling examples of less frequent classes. In order to create a system that can be trained to rely more explicitly on antecedent information, we have designed a neural network classifier for our task.

Artificial neural networks are networks of classifiers, usually organised into layers, where the outputs of the classifiers in one layer are fed as inputs to the classifiers of the next layer. The individual classifier cells map a vector of inputs to a single output with a non-linear function parametrised with a set of weights similar to the weights of a maximum entropy classifier. A DP algorithm by the name of backpropagation (Rumelhart et al., 1986) allows computing the gradients of an error function of the network outputs with respect to all the weights in the network in polynomial time, so the network can be trained efficiently with a variant of the gradient descent algorithm.

The main advantage of a neural network over a single classifier is that it is capable of learning and representing latent variables. The classifiers in the hidden layers of the network, whose outputs do not correspond directly to

network outputs, but are connected to the inputs of another layer of classifiers, can learn to recognise abstract features of the input data that are then made available to the next layer. Since the gradients of the parameters of the hidden layers are computed with backpropagation based on an error function involving only the predictions of the final output layer, no supervision for the intermediate abstract representation is required.

Neural networks have recently gained some popularity in natural language processing. They have been applied to tasks such as language modelling (Bengio et al., 2003; Schwenk, 2007), translation modelling in statistical machine translation (Le et al., 2012), but also part-of-speech tagging, chunking, named entity recognition and semantic role labelling (Collobert et al., 2011). In tasks related to anaphora resolution, standard feed-forward neural networks have been tested as a classifier in an anaphora resolution system (Stuckardt, 2007), but the idea of using a neural network for cross-lingual pronoun prediction is novel in our work.

In the case of our pronoun prediction network, the introduction of a hidden layer should enable the classifier to learn abstract concepts such as gender and number that are useful across multiple output categories, so that the performance of sparsely represented classes can benefit from the training examples of the more frequent classes. Additionally, as we shall see in Section 8.5, the neural network's capacity for dealing with latent variables allows us to represent the links between anaphoric pronouns and their antecedents as latent variables, dispensing with the need for a separately trained anaphora resolution system.

The overall structure of the network is shown in Fig. 8.3. As inputs, it takes the same features that were available to the baseline ME classifier, based on the source pronoun (**P**) with three words of context to its left (**L1** to **L3**) and three words to its right (**R1** to **R3**) as well as the words aligned to the syntactic head words of all possible antecedent candidates as found by BART (**A**). All words are encoded as one-hot vectors whose dimensionality is equal to the vocabulary size. If multiple words are aligned to the syntactic head of an antecedent candidate, their word vectors are averaged with uniform weights. The resulting vectors for each antecedent are then averaged with weights defined by the posterior distribution of the anaphora resolver in BART (p_1 to p_3; see also Fig. 8.2).

The network has two hidden layers. The first layer (**E**) maps the input word vectors to a low-dimensional representation. In this layer, the embedding weights for all the source language vectors (the pronoun and its 6 context words) are tied, so if two words are the same, they are mapped to the same lower-dimensional embedding regardless of their position relative to the pronoun. The embedding of the antecedent word vectors is independent, as these word vectors represent target language words. The entire embedding layer is then mapped to another hidden layer (**H**), which is in turn connected to a softmax output layer (**S**) with 6 outputs representing the classes *ce*, *elle*, *elles*, *il*,

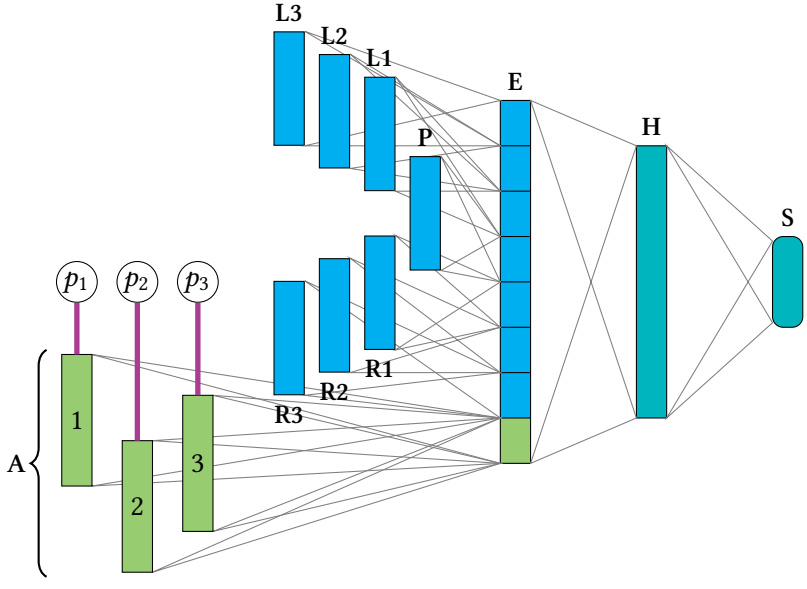

Figure 8.3. Neural network for pronoun prediction

ils and OTHER. The softmax layer estimates a normalised probability distribution over the different outputs. The non-linearity of both hidden layers is the logistic sigmoid function, $f(x) = 1/(1+e^{-x})$. We obtained similar results (not detailed here) with the hyperbolic tangent transfer function, $f(x) = \tanh x$, and with rectified linear units whose transfer function is $f(x) = \max(0, x)$.

In all experiments reported in this chapter, the dimensionality of the source and target language word embeddings is 20, resulting in a total embedding layer size of 160, and the size of the last hidden layer is equal to 50. These sizes are very small. In experiments with larger layer sizes, we obtained similar, but no better results.

The neural network is trained with minibatch stochastic gradient descent with backpropagated gradients using the RMSPROP algorithm (Algorithm 2).[2] The algorithm repeatedly samples a small set or minibatch $\mathbf{M}$ of training examples from the training corpus and computes the gradients G of the objective function with respect to the network parameters. It then tries to improve the value of the objective function by applying a small correction to the parameter vector. The magnitude of the correction depends, among other things, on the learning rate α, and its direction is a function of the gradients of the current iteration and the gradients seen in previous iterations.

The objective function that we optimise for is cross-entropy, the standard error function for neural networks with softmax output layers. For a single

[2]Our training procedure is greatly inspired by a series of on-line lectures held by Geoffrey Hinton in 2012 (`https://www.coursera.org/course/neuralnets`, 10 September 2013).

training example, it is computed as

$$E = -\sum_i t_i \log y_i, \tag{8.1}$$

where the sum is over the units of the output layer representing the output classes, t_i is the target value found in the training set and y_i is the probability assigned to this class by the neural network with the current weights. `Fprop` and `Bprop` are functions implementing the forward and backward propagation pass through the network, respectively.

In contrast to standard gradient descent, RMSPROP normalises the magnitude of the gradient components by dividing them by a root-mean-square moving average accumulated in the vector R (lines 10 and 11). We find that this leads to faster convergence. We also apply some other heuristics to improve the speed of convergence. In most cases, there is no principled justification for the numerical values of the parameters of these heuristics, but they are fixed empirically to improve the observed time required to achieve convergence when training our network or earlier versions of it.

- Momentum is used to even out gradient oscillations, so the direction of the weight adjustment made in each iteration of the optimisation procedure, which is stored in the vector Δ, is equal to m times the direction of the previous iteration plus the contribution of the current iteration (line 12). The momentum parameter m is set to the constant 0.9 in all our experiments.
- The global learning rate is multiplied with a gain factor Γ_i for each individual weight (line 12). Initially set to 1, the gain factor is increased by adding 0.05 whenever the gradient of a weight has the same sign in two subsequent minibatch iterations. When the gradient changes sign, the gain factor is decreased by multiplying with 0.95 (lines 14–20).
- The global learning rate is adjusted according to training progress. Let d be the number of times the training error decreased in the last 6 epochs. If $d < 4$, i. e., if the training error increased more than twice, then the learning rate is decreased by 20 %. Otherwise, it is increased by 5 % stochastically after each epoch with probability $0.3d/6$. After each adjustment, the learning rate is held constant for at least 6 epochs (lines 22–31).

Good settings of the initial learning rate and the weight cost parameter (both around 0.001 in most experiments), as well as other training parameters, were found by manual experimentation. The initial learning rate is set to the highest value that reliably leads to convergence. The weight cost parameter is selected to minimise validation error. Generally, we train our networks for 300 epochs, which seems to be amply sufficient for the network to converge. We compute the validation error on a held-out set of some 10 % of the training data after each epoch and use the set of parameters that achieves the lowest validation error for testing.

Algorithm 2 RMSPROP neural network training algorithm

Input: training set $\mathbf{T}$, learning rate α, number of epochs e, minibatch size b, momentum parameter m, start weights W, a validation set
Output: optimised weights

for all weight components i **do**
  $R_i \leftarrow 1; \quad \Gamma_i \leftarrow 1; \quad \Delta_i \leftarrow 0$
end for
$E_{\text{best}} \leftarrow \infty$
for $i \leftarrow 1$ **to** e **do**
  for $j \leftarrow 1$ **to** $\texttt{Size(T)}/b$ **do**
    $\mathbf{M} \leftarrow b$ examples from $\mathbf{T}$, sampled without replacement
    $y \leftarrow \texttt{Fprop(M,}W\texttt{)}$
    $G \leftarrow \texttt{Bprop(M,}W\texttt{,}y\texttt{)}$
    $R \leftarrow 0.9R + 0.1G^2$
    $G' \leftarrow G/\sqrt{R}$
    $\Delta \leftarrow m\Delta - \alpha\Gamma G'$
    $W \leftarrow W + \Delta$
    for all weight components i **do**
      if G_i has the same sign as for the last minibatch **then**
        $\Gamma_i \leftarrow \Gamma_i + 0.05$
      else
        $\Gamma_i \leftarrow 0.95\Gamma_i$
      end if
    end for
  end for
  if $c > 5$ **then**
    $d \leftarrow$ number of times training error decreased in the last 6 epochs
    if $d < 4$ **then**
      $\alpha \leftarrow 0.8\alpha$
    else
      $\alpha \leftarrow 1.05\alpha$ with probability $0.3d/6$
    end if
    $c \leftarrow 0$
  end if
  $c \leftarrow c + 1$
  $E_{\text{val}} \leftarrow$ error on validation set
  if $E_{\text{val}} < E_{\text{best}}$ **then**
    $W_{\text{best}} \leftarrow W$
    $E_{\text{best}} \leftarrow E_{\text{val}}$
  end if
end for
return W_{best}

All vector operations are performed elementwise.

Table 8.5. *Neural network classifier with pronouns resolved by BART*

	TED (Accuracy: 0.700)			News commentary (Accuracy: 0.576)		
	P	R	F	P	R	F
ce	0.634	0.747	0.686	0.477	0.344	0.400
elle	0.756	0.617	0.679	0.498	0.401	0.444
elles	0.679	0.319	0.434	0.565	0.116	0.193
il	0.719	0.591	0.649	0.655	0.626	0.640
ils	0.663	0.940	0.778	0.570	0.834	0.677
OTHER	0.743	0.678	0.709	0.567	0.573	0.570

Since the source context features are very informative and it is comparatively more difficult to learn from the antecedents, the network sometimes has a tendency to overfit to the source features and ignore the information coming from the antecedents. This problem can be solved effectively by removing the source features from a part of the training material, forcing the network to learn from the information contained in the antecedents. In all experiments in this paper, we zero out each individual source feature (input layers **P**, **L1** to **L3** and **R1** to **R3**) stochastically with a probability of 50 % every time a training example is presented to the network. At test time, no information is zeroed out.

Classification results with this network are shown in Table 8.5. The accuracy increases slightly for the TED test set and remains exactly the same for the News commentary corpus. However, a closer look on the results for individual classes reveals that the neural network makes better predictions for almost all classes. In terms of F-score, the only class that becomes slightly worse is the OTHER class for the News commentary corpus because of lower recall, indicating that the neural network classifier is less biased towards using the uninformative OTHER category. Recall for *elle* and *elles* increases considerably, but especially for *elles* it is still quite low. For the TED data, the increase in recall comes with some loss in precision, but the net effect on F-score is clearly positive.

8.5 Latent Anaphora Resolution

Considering Fig. 8.1 again, we note that the bilingual setting of our classification task adds some information not available to the monolingual anaphora resolver that can be helpful when determining the correct antecedent for a given pronoun. Knowing the gender of the translation of a pronoun limits the set of possible antecedents to those whose translation is morphologically

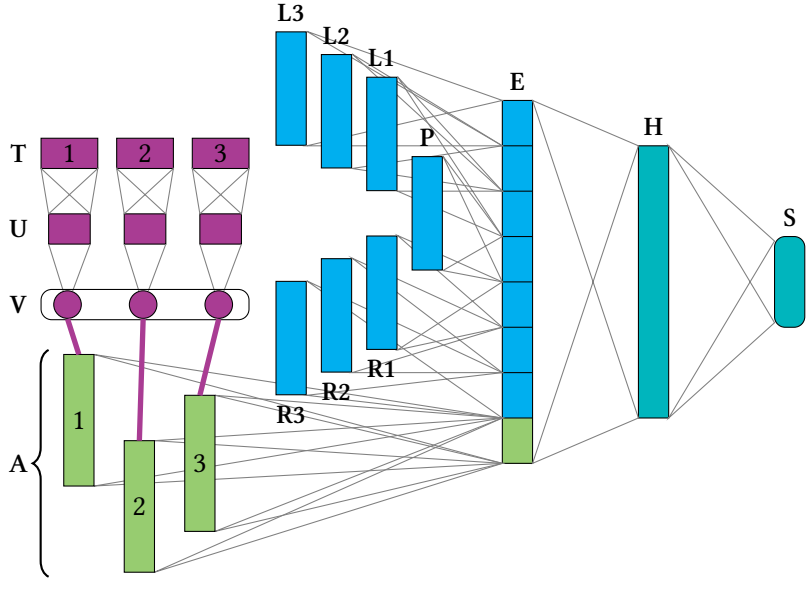

Figure 8.4. Neural network with latent anaphora resolution

compatible with the target language pronoun. Exploiting this fact and the capacity of neural networks for learning hidden representations gives us the possibility to treat the anaphoric links as latent variables, which allows us to avoid the use of data manually annotated for coreference, in line with the modelling assumptions we have chosen to adopt for this thesis (Section 1.3).

To achieve this, we extend the network with a component to predict the probability of each antecedent candidate to be the correct antecedent (Fig. 8.4). The extended network is identical to the previous version except for the upper left part dealing with anaphoric link features. The only difference between the two networks is the fact that anaphora resolution is now performed by a part of our neural network itself instead of being done by an external module and provided to the classifier as an input.

In this setup, we still use some parts of the BART toolkit to extract markables and compute features. However, we do not make use of the machine learning component in BART that makes the actual predictions. Since this is the only component trained on coreference-annotated data in a typical BART configuration, no coreference annotations are used anywhere in our system even though we continue to rely on the external anaphora resolver for preprocessing to avoid implementing our own markable and feature extractors and to make comparison easier.

For each candidate markable identified by BART's preprocessing pipeline, the anaphora resolution model receives as input a link feature vector (**T**) describing relevant aspects of the antecedent candidate-anaphora pair. This

feature vector is generated by the feature extraction machinery in BART and includes a standard set of features for coreference resolution. We use the following feature extractors in BART, each of which can generate multiple features:

- *Anaphor mention type:* Checks whether the anaphor is a proper name, a noun phrase or a pronoun, and if so what type of pronoun.
- *Gender match:* Checks whether the anaphor and the antecedent agree in gender.
- *Number match:* Checks whether the anaphor and the antecedent agree in number.
- *String match:* Checks for a string match between the anaphor and the antecedent.
- *Alias feature:* Checks for fuzzy matches between the anaphor and the antecedent with the help of some heuristics (see Soon et al., 2001).
- *Appositive position feature:* Checks whether the anaphor could be an apposition of the antecedent (see Soon et al., 2001).
- *Semantic class:* Encodes the semantic class of the anaphor (see Soon et al., 2001).
- *Semantic class match:* Checks whether the semantic classes of the anaphor and the antecedent match.
- *Binary distance features:* Encode whether the anaphor and the antecedent are in the same or in adjacent sentences.
- *First mention:* Encodes whether the antecedent is the first mention in a sentence.

Our baseline set of features is borrowed wholesale from a working coreference system. It is based on the elementary feature set of Soon et al. (2001) with some additional features from work by Uryupina (2006). Many of the features such as those indicating that the anaphor is a pronoun or that it is not a named entity are not relevant to the pronoun prediction task. To ensure that the features used by our network are exactly the same as those used by BART, we do not manipulate the feature extractor list at this point. Instead, we remove all features that assume constant values in the training set when resolving antecedents for the set of pronouns we consider. Ultimately, we are left with a basic set of 37 anaphoric link features that are fed as inputs to our network. These features are exactly the same as those available to the anaphora resolution classifier in the BART system used in the previous section.

Each training example for our network can have an arbitrary number of antecedent candidates, each of which is described by an antecedent word vector (**A**) and by an anaphoric link vector (**T**). The anaphoric link features are first mapped to a regular hidden layer with logistic sigmoid units (**U**). The activations of the hidden units are then mapped to a single value, which

functions as an element in a softmax layer over all antecedent candidates (**V**). This softmax layer assigns a probability to each antecedent candidate, which we then use to compute a weighted average over the antecedent word vector, replacing the probabilities p_i in Fig. 8.2 and Fig. 8.3.

At training time, the network's anaphora resolution component is trained in exactly the same way as the rest of the network. The error signal from the embedding layer is backpropagated both to the weight matrix defining the antecedent word embedding and to the anaphora resolution subnetwork. Note that the number of weights in the network is the same for all training examples even though the number of antecedent candidates varies because all weights related to antecedent word features and anaphoric link features are shared between all antecedent candidates.

One slightly uncommon feature of our neural network is that it contains an internal softmax layer (**V**) to generate probabilities normalised over all possible antecedent candidates. Moreover, weights are shared between all antecedent candidates, so the inputs of our internal softmax layer share dependencies on the same weight variables. When computing derivatives with backpropagation, these shared dependencies must be taken into account. In particular, the outputs y_i of the antecedent resolution layer are the result of a softmax applied to functions of some shared variables $q_1, \ldots, q_n$:

$$y_i = \frac{\exp f_i(q_1, \ldots, q_n)}{\sum_k \exp f_k(q_1, \ldots, q_n)} \tag{8.2}$$

The derivatives of any y_i with respect to a q_j, which can be any of the weights in the anaphora resolution subnetwork, have dependencies on the derivatives of the other softmax inputs with respect to q_j:

$$\frac{\partial y_i}{\partial q_j} = y_i \left(\frac{\partial f_i(q_1, \ldots, q_n)}{\partial q_j} - \sum_k y_k \frac{\partial f_k(q_1, \ldots, q_n)}{\partial q_j} \right) \tag{8.3}$$

This makes the implementation of backpropagation for this part of the network somewhat more complicated, but it has no significant impact on training time.

Experimental results for this network are shown in Table 8.6. Compared with Table 8.5, we note that the overall accuracy is only very slightly lower for TED, and for the News commentaries it is actually better. When it comes to F-scores, the performance for *elles* improves, while the effect on the other classes is a bit more mixed. Even where it gets worse, the differences are not dramatic considering that we have eliminated the manually annotated coreference training set, a very knowledge-rich resource, from the training process. This demonstrates that it is possible, in our classification task, to obtain good results without using any data manually annotated for anaphora and to rely entirely on unsupervised latent anaphora resolution.

Table 8.6. *Neural network classifier with latent anaphora resolution*

	TED (Accuracy: 0.696)			News commentary (Accuracy: 0.597)		
	P	R	F	P	R	F
ce	0.618	0.722	0.666	0.419	0.368	0.392
elle	0.754	0.548	0.635	0.547	0.460	0.500
elles	0.737	0.340	0.465	0.539	0.135	0.215
il	0.718	0.629	0.670	0.623	0.719	0.667
ils	0.652	0.916	0.761	0.596	0.783	0.677
OTHER	0.741	0.682	0.711	0.614	0.544	0.577

8.6 Further Improvements

The results presented in the preceding section represent a clear improvement over the ME classifiers in Table 8.4, even though the overall accuracy increases only slightly. Not only does our neural network classifier achieve better results on the classification task at hand without requiring an anaphora resolution classifier trained on manually annotated data, but it performs clearly better for the feminine categories that reflect minority choices requiring knowledge about the antecedents. Nevertheless, the performance is still not entirely satisfactory.

By subjecting the output of our classifier on a development set to a manual error analysis, we found that a fairly large number of errors belong to two error types: On the one hand, the preprocessing pipeline used to identify antecedent candidates does not always include the correct antecedent in the set presented to the neural network. Whenever this occurs, it is obvious that the classifier cannot possibly find the correct antecedent. Out of 76 examples of the category *elles* that had been mistakenly predicted as *ils*, we found that 43 suffered from this problem. In other classes, the problem seems to be somewhat less common, but it still exists. On the other hand, in many cases (23 out of 76 for the category mentioned before) the anaphora resolution subnetwork does assign the highest probability to an antecedent which, even if possibly incorrect, belongs to the right gender/number group, but it still predicts an incorrect pronoun. This may indicate that the network has difficulties learning a correct gender/number representation for all words in the vocabulary.

8.6.1 Relaxing Markable Extraction

The pipeline we use to extract potential antecedent candidates is borrowed from the BART anaphora resolution toolkit. BART uses a syntactic parser to identify noun phrases as markables. When extracting antecedent candid-

ates for coreference prediction, it starts by considering a window consisting of the sentence in which the anaphoric pronoun is located and the two immediately preceding sentences. Markables in this window are checked for morphological compatibility in terms of gender and number with the anaphoric pronoun, and only compatible markables are extracted as antecedent candidates. If no compatible markables are found in the initial window, the window is successively enlarged one sentence at a time until at least one suitable markable is found.

Our error analysis shows that this procedure misses some relevant markables for at least two reasons. On the one hand, the initial three-sentence extraction window is too small. On the other hand, the morphological compatibility check incorrectly filters away some markables that should have been considered as candidates. By contrast, the extraction procedure does extract quite a number of first and second person noun phrases (*I, we, you* and their oblique forms) in the TED talks, which are extremely unlikely to be the antecedent of a later occurrence of *he, she, it* or *they*. As a first step, we therefore adjust the extraction criteria to our task by increasing the initial extraction window to six sentences, excluding first and second person markables and removing the morphological compatibility requirement. The compatibility check is still used to control expansion of the extraction window, but it is no longer applied to filter the extracted markables. This increases the accuracy to 0.701 for TED and 0.602 for the News commentaries, while the performance for *elles* improves to F-scores of 0.531 (TED; P 0.690, R 0.432) and 0.304 (News commentaries; P 0.444, R 0.231), respectively. Note that these and all the following results are not directly comparable to the ME baseline results in Table 8.4, since they include modifications and improvements to the training data extraction procedure that might possibly lead to benefits in the ME setting as well.

8.6.2 Adding Lexicon Knowledge

In order to make it easier for the classifier to identify the gender and number properties of infrequent words, we extend the word vectors with features indicating possible morphological features for each word. In early experiments with ME classifiers, we found that our attempts to do proper gender and number tagging in French text did not improve classification performance noticeably, presumably because the annotation was too noisy. In more recent experiments, we just add features indicating all possible morphological interpretations of each word, rather than trying to disambiguate them. To do this, we look up the morphological annotations of the French words in the Lefff dictionary (Sagot et al., 2006) and introduce a set of new binary features to indicate whether a reading of a word with a particular set of morphosyntactic properties occurs in that dictionary. These binary features are then

Table 8.7. *Final classifier results*

	TED (Accuracy: 0.713)			News commentary (Accuracy: 0.626)		
	P	R	F	P	R	F
ce	0.611	0.723	0.662	0.492	0.324	0.391
elle	0.749	0.596	0.664	0.526	0.439	0.478
elles	0.602	0.616	0.609	0.547	0.558	0.552
il	0.733	0.638	0.682	0.599	0.757	0.669
ils	0.710	0.884	0.788	0.671	0.878	0.761
OTHER	0.760	0.704	0.731	0.681	0.526	0.594

added to the one-hot representation of the antecedent words. Doing so improves the classifier accuracy to 0.711 (TED) and 0.604 (News commentaries), while the F-scores for *elles* reach 0.589 (TED; P 0.649, R 0.539) and 0.500 (News commentaries; P 0.545, R 0.462), respectively.

8.6.3 More Anaphoric Link Features

Even though the modified antecedent candidate extraction with its larger context window and without the morphological filter results in better performance on both test sets, additional error analysis reveals that the classifier has greater problems identifying the correct markable in this setting. One reason for this may be that the baseline anaphoric link feature set described above (Section 8.5) only includes two very rough binary distance features which indicate whether or not the anaphora and the antecedent candidate occur in the same or in immediately adjacent sentences. With the larger context window, this may be too unspecific. In our final experiment, we therefore enable some additional features which are implemented in BART, but disabled in the baseline system:

- Distance in number of markables
- Distance in number of sentences
- Sentence distance, log-transformed
- Distance in number of words
- Part of speech of head word

Most of these encode the distance between the anaphora and the antecedent candidate in more precise ways. Complete results for this final system are presented in Table 8.7.

Including these additional features leads to another slight increase in accuracy for both corpora, with similar or increased classifier F-scores for most classes except *elle* in the News commentary experiment. In particular, we

should point out the performance of our benchmark classifier for *elles*, which suffered from extremely low recall in the first classifiers and approaches the performance of the other classes, with nearly balanced precision and recall, in this final system. Since *elles* is a low-frequency class and cannot be reliably predicted using source context alone, we interpret this as evidence that our final neural network classifier has incorporated some relevant knowledge about pronominal anaphora that the baseline ME classifier and earlier versions of our network have no access to. This is particularly remarkable because no data manually annotated for coreference was used for training.

8.7 Conclusion

In this chapter, we have introduced cross-lingual pronoun prediction as an independent natural language processing task. Even though it is not an end-to-end task, pronoun prediction is interesting for several reasons, not least because of its relation to pronoun translation in SMT. We have shown that pronoun prediction can be effectively modelled in a neural network architecture with relatively simple features. More importantly, we have demonstrated that the task can be exploited to train a classifier with a latent representation of anaphoric links. With parallel text as its only supervision this classifier achieves a level of performance that is similar to, if not better than, that of a classifier using a regular anaphora resolution system trained with manually annotated data.

9. Pronoun Prediction in SMT

The pronoun prediction model developed in the previous chapter maps pronoun translations to a probability score given information extracted from a piece of bilingual context potentially covering multiple sentences. Such a model can easily be integrated in the document-level decoding framework presented in the first part of this thesis. This chapter concludes our experimental work by combining the different components we have developed into one system, a document-level SMT system built around the Docent decoder with a neural network model for pronoun prediction. We study some of the difficulties that arise when the pronoun prediction model is used in an SMT setting and investigate the output of the enhanced system with the help of both automatic methods and a targeted manual evaluation experiment.

9.1 Integrating the Anaphora Model into Docent

Predicting the correct translation of an anaphoric pronoun has two parts, identifying its antecedent in the source language and finding out what linguistic elements best represent the input pronoun in the target language, also taking into account the translation of the antecedent. The neural network model of the previous chapter (Fig. 8.4, p. 128) incorporates anaphora resolution and target element selection in a single neural network classifier. It is trained with backpropagation on training examples extracted from word-aligned parallel text, and the anaphoric links are treated as latent variables. The inputs of the neural network consist of anaphora context features and anaphoric link features, which are extracted from the source language part of the training and test examples, and antecedent features, which are extracted from the translation.

The SMT decoder takes source language material as input and generates a translation. At decoding time, only the features depending on the translated output are variable as the translation is generated and updated in the decoding process. Features derived from the input are fixed. In particular, the part of the network that deals with anaphoric links (layers **T**, **U** and **V** in Fig. 8.4) are independent of the translation and can be precomputed. Instead of implementing the anaphora resolution component of the network as a part of the SMT decoder, we therefore run it as a preprocessing step and integrate it into the coreference resolution toolkit BART (Versley et al., 2008), which we also use to extract markables and anaphoric link features. In BART, the

neural network simply replaces the standard markable ranking component. Before running the SMT decoder, we process the source file with this modified version of BART to extract markables and compute the probabilities of network layer **V** as input for the translation step. Thus, the anaphora resolution subnetwork, which was united with the pronoun prediction classifier at training time, is now again run separately at decoding time.

The remaining parts of the neural network are added as a feature function to the Docent document-level SMT decoder. At the beginning of a decoding run, the feature module identifies all relevant anaphoric pronouns in the document according to some filter criteria. In our English–French experiments, we consider all occurrences of the English pronouns *it* and *they*. The module also identifies all the markables the anaphora resolver recognised as antecedent candidates for one of the target anaphors with a probability exceeding some small threshold value. The purpose of the threshold is to avoid spending an inordinate amount of time on numerous low-probability candidates. It is set to 0.01 in our experiments.

The markables passing these filters are stored in a data structure that links anaphors to their antecedent candidates as well as antecedent markables to potential anaphors. Then, the document is scored by extracting all necessary information from the markables and making a forward propagation pass through the neural network. The feature score of a single anaphor is the logarithm of the probability assigned by the softmax output layer **S** to the pronoun translation found in the current document state. These scores are summed over the complete document and cached in the data structure representing the anaphor.

Whenever the document state is modified, the decoder identifies the anaphoric pronouns and antecedent candidates affected by the modification. It then recomputes and updates the scores of those anaphors which are affected by the modification themselves or whose antecedent candidates are affected by it.

9.2 Weakening Prior Assumptions in the SMT Models

Even though the standard translation and language models of phrase-based SMT do not model pronominal anaphora explicitly, they make strong prior assumptions about how pronouns should be translated. Table 9.1 shows the top ten translations of the single-word phrases *it* and *they* in a phrase table created with the WMT 2014 English–French training data. The entries are ordered by the geometric mean of the probability of the target phrase given the source and that of the source phrase given the target, equivalent to a log-linear combination with equal weights.

In both singular and plural, the obvious translation equivalents *il* and *elle* or *ils* and *elles* top the lists. Two-word phrases with the conjunction *que*

Table 9.1. *Top ten translations of* it *and* they *in an English–French phrase table*

		$p(t\|s)$	$p(s\|t)$	$\bar{p}_{\text{geom}}$			$p(t\|s)$	$p(s\|t)$	$\bar{p}_{\text{geom}}$
it	il	0.258	0.379	0.312	they	ils	0.234	0.543	0.357
	elle	0.100	0.409	0.202		elles	0.109	0.432	0.217
	qu' il	0.043	0.137	0.077		qu' ils	0.081	0.296	0.155
	c'	0.018	0.293	0.073		qu' elles	0.031	0.258	0.090
	qu' elle	0.023	0.186	0.065		ils ont	0.016	0.233	0.060
	cela	0.012	0.156	0.044		leur	0.039	0.030	0.034
	lui	0.014	0.111	0.039		ceux-ci	0.007	0.160	0.033
	celle-ci	0.005	0.168	0.030		celles-ci	0.005	0.131	0.025
	on	0.013	0.058	0.028		, ils	0.008	0.082	0.025
	, il	0.014	0.043	0.025		Ils	0.007	0.043	0.017

follow closely, reflecting the fact that the English complementiser *that* can frequently be omitted in places where French requires the use of *que*. In the singular, the pronoun *c'* of the construction *c'est*, translating into *it is*, and the demonstrative pronoun *cela* also achieve high scores. All of this is entirely unsurprising and intuitive, and the translations contained in the phrase table supply translational equivalents for many frequent uses of the English pronouns. While there is little semantic difference between the various translations, the correct choice between them is governed by all manner of linguistic constraints, ranging from the syntactic relations that control the choice between subject forms like *il* and object forms like *lui* to the discourse mechanisms that may trigger the use of *cela* instead. The translation model has no notion of these constraints, but it assigns vastly different scores to different translation alternatives based on their frequency in the training corpus. Thus, all other things being equal, the decoder will always prefer masculine translations over feminine ones, and given the choice between *il est* and *c'est* as translations of *it is*, which is often largely a matter of style, the former will be preferred. These preferences may be overridden by the immediately surrounding context, which may induce the use of a multi-word phrase with different top-scoring translations or cause the language model to score up another translation, but none of these dependencies works over a longer distance than a handful of words.

Sometimes the *n*-gram language model has amazing ways of selecting pronouns without actually knowing anything about anaphora. Consider the following example:

(9.1) a. *Input:* It is necessary to say that the car insurance is something important, not only because *it* covers the driver over possible wrecks, but because *it* represents an important cost [...] (*news-test2008*)

b. *Reference translation:* Il faut dire que l'assurance automobile est quelque chose d'important, non seulement parce qu'*elle* couvre le conduc-

teur face à d'éventuels sinistres, mais aussi parce qu'*elle* représente une importante dépense, [...]

c. *Baseline MT output:* Il est nécessaire de dire que l'assurance automobile est quelque chose d'important, non seulement parce qu'*elle* couvre le conducteur au possible des épaves, mais parce qu'*il* représente un coût important, [...]

The MT output is generated by a baseline Moses system trained on a substantial part of the WMT 2014 parallel training data which has a 6-gram language model trained on news text from the News commentary corpus and the News crawl corpus provided by the shared task organisers and the French Gigaword corpus from LDC. The system does not have any models specifically dealing with pronominal anaphora. Nevertheless, the first instance of the pronoun *it* referring to *car insurance* is correctly rendered with the French feminine pronoun *elle*, although its French antecedent, the feminine noun phrase *l'assurance automobile*, is far beyond the history of the 6-gram language model. It turns out that the *n*-gram context of the anaphoric pronoun is highly predictive of the identity of its antecedent. The language model training corpus contains the following two sentences, both of which overlap with the test sentence in the 5-gram *qu'elle couvre le conducteur*:

(9.2) a. Paradoxalement, cette garantie n'est pas toujours incluse dans l'assurance auto bien qu'elle couvre le conducteur, qu'il soit propriétaire du véhicule ou non.

b. La réponse à ces questions tout autant que les garanties liées à l'« individuelle conducteur », qui n'est pas toujours incluse dans l'assurance auto bien qu'elle couvre le conducteur, permet de différencier deux contrats.

In both cases, the antecedent of *elle* is the noun phrase *l'assurance auto*, a shortened form of the phrase *l'assurance automobile* of the previous example with the same morphosyntactic features. No corresponding sentence with the masculine pronoun *il* occurs in the training set. Far from demonstrating any capability of handling anaphoric pronouns, our example illustrates the *n*-gram model's astonishing capacity for acquiring world knowledge. What has really been learnt is the gender of the noun phrase that a pronoun in the given context *typically* refers to rather than the gender of the antecedent it *specifically* refers to in this example.

If the goal of running an SMT system is to create the best translations possible given the current state of the art, then it is a useful strategy to exploit the skewness of the pronoun frequency distribution to make good, if uninformed, guesses in as many cases as possible. However, since our goal is to develop better pronoun models, the effect of the frequency priors is undesired because it distorts the real performance of the pronoun models. If the guesswork of the language and translation models leads to the right conclusion,

it may disguise mistakes of the anaphora model and generate correct output in spite of it. Conversely, if the language and translation models impose a more frequent pronoun choice despite better advice of the anaphora model, spurious errors are introduced.

While this is a problem that arises because the translation and language models incompetently interfere with the work of the anaphora model, the pronoun prediction model also interferes with some choices that the core SMT models are better at solving. The pronoun classifier of the previous chapter focuses on the French pronouns *il*, *elle*, *ils* and *elles* when they occur as translations of an English third-person subject pronoun. All other translations of the English pronouns are lumped together into a single class OTHER. In concrete decoding situations, however, the class OTHER as such never occurs; instead, the model is confronted with the problem of distributing the probability mass reserved for this class to a large variety of different candidate translations, similar to the way the *n*-gram language model must distribute the total probability mass reserved for the class of unseen words to individual, and possibly competing, instances of unseen words. Unless the model were extended with some kind of language modelling capacity, this distribution would be arbitrary because, having established that a candidate translation does not contain any of the pronouns it knows about, the anaphora model has no useful information to score it.

Luckily, both of these difficulties can be overcome at once by decoupling the anaphora model from the translation and language models and letting each model do what it knows most about. The key is to remove the OTHER category from the pronoun prediction model and to remove information about the identity of the pronouns it models from the language and translation models. We replace each occurrence of the pronouns *il* and *elle* that is aligned to an English *it* and each occurrence of *ils* and *elles* that is aligned to an English *they* by a placeholder while training the language model and the translation model. Since the language model needs information about some features of the pronouns to fit them correctly into the surrounding context, we use four different placeholders for capitalised versus lowercased and singular versus plural pronouns, respectively. Thus, we replace *il* by a placeholder called LCPRONOUN-SG and *Elles* by a placeholder called UCPRONOUN-PL if they are aligned to *it* or *they*, respectively. The scores of the translation model and the language model are computed over the text with these placeholders. We use two copies of the pronoun prediction model in the decoder, one to handle singular pronouns and one to handle plural pronouns. If a target phrase contains a placeholder, separate hypotheses with all compatible pronouns are generated and scored by the appropriate pronoun prediction model. If no placeholder occurs in the translation, the pronoun predictors do not add any score.

With this model, the target language pronouns *il*, *elle*, *ils* and *elles* can be generated in two ways. In the generation path we are primarily inter-

ested in, the translation model generates a pronoun placeholder. The translation model and the language model calculate their scores based on the placeholder. Then, a pronoun is generated from the placeholder, and the pronominal anaphora model calculates its score based on the pronoun. This happens whenever the target pronoun is aligned to *it* or *they* in the source language. In the second generation path, a pronoun is generated directly by the translation model with a phrase pair in which the source pronouns *it* or *they* either do not occur or are not aligned to the target pronoun. In this case, the translation and language models get to see the pronoun itself instead of a placeholder, and the pronoun prediction model is not active at all. Thus, the translation model and the language model contain both pronoun placeholders and concrete instances of pronouns.

For the translation model, this does not pose any difficulties at training time. Since the model is trained on word-aligned parallel text, it is easy to check whether a given pronoun instance is aligned to one of the English source pronouns and to insert a placeholder only if this is the case. For a language model trained on monolingual data, this will not work because there is no English source text, so it is not trivial to find out whether or not a target language pronoun would be aligned to *it* or *they* in a hypothetical source language text. To train a model with an approximately correct distribution of pronouns and placeholders, we first create a 6-gram language model over the target language side of a part of the translation model training corpus with placeholders inserted according to the aligned source words. Next, we use this model to insert placeholders into the actual training corpus by running the Viterbi decoder for n-gram-based disambiguation included in the SRILM language modelling toolkit (Stolcke et al., 2011). Finally, we train a 6-gram language model on this artificially annotated training corpus and use it in our SMT system.

9.3 SMT Experiments

To test our anaphora model, we run a series of experiments integrating the model into phrase-based English–French SMT systems for the two text types we tested our classifiers on in the previous chapter. The systems incorporate the document-level anaphora model in the local search decoder developed in the first part of this thesis.

9.3.1 Baseline Systems

Decoding is done in two steps. First, we run a sentence-level phrase-based SMT system with the Moses decoder (Koehn et al., 2007). The output of this decoder is then used to initialise the Docent local search decoder described in Chapter 4. At the same time, we use it as a baseline.

The fundamental setup of our baseline system for News data is loosely based on the system submitted by Cho et al. (2013) to the WMT 2013 shared task. Our phrase table is trained on data taken from the News commentary, Europarl, UN, Common crawl and 10^9 corpora. The first three of these corpora were included integrally into the training set after filtering out sentences of more than 80 words. The Common crawl and 10^9 data sets were run through an additional filtering step with an SVM classifier, closely following Mediani et al. (2011). The phrase table of the baseline system is the same as that of the document-level system and is created by reinserting pronouns in a phrase table with placeholders as described in the previous section. At each occurrence of the placeholders LCPRONOUN-SG, LCPRONOUN-PL, UCPRONOUN-SG and UCPRONOUN-PL, the applicable pronouns are inserted with equal probabilities. As a result, the choice between these pronouns is entirely left to the language model in the baseline system.

The system includes three language models, a regular 6-gram model with modified Kneser-Ney smoothing (Chen and Goodman, 1998) trained with KenLM (Heafield, 2011), a 4-gram bilingual language model (Niehues et al., 2011) with Kneser-Ney smoothing trained with KenLM and a 9-gram model over Brown clusters (Brown et al., 1992) with Witten-Bell smoothing (Witten and Bell, 1991) trained with SRILM (Stolcke et al., 2011). In addition to the three language models, the baseline system uses the standard set of features for phrase-based SMT with four phrase table scores, a phrase penalty, a word penalty, an out-of-vocabulary penalty and a geometric distortion model. No lexical reordering models are included.

The TED system is identical to the News system, but the TED parallel training corpus from the WIT^3 distribution (Cettolo et al., 2012) is added to the translation model training set, and the monolingual French WIT^3 training data is added to the LM corpus. The feature weights of the baseline systems are optimised with MERT (Och, 2003) against the *newstest2011* and the *dev2010* development set for the News and the TED system, respectively.

9.3.2 Document-Level Decoding with Anaphora Models

In the document-level decoder, the anaphora model is added to the baseline configuration in the form of two extra feature functions. Each of the feature functions corresponds to a separate instance of the neural network classifier. One of them handles the singular pronoun *it* and makes a binary choice between *il* and *elle*, and the other handles the plural pronoun *they* and makes a binary choice between *ils* and *elles*. Examples where *it* is aligned to *ils* or *elles* or where *they* is aligned to *il* or *elle* are not handled by the anaphora model. The anaphora feature functions are only active if the input pronoun *it* or *they* is aligned to a pronoun placeholder on the target side. If there is no placeholder corresponding to a specific input pronoun, the anaphora models

Table 9.2. *Neural network configurations and intrinsic performance*

	$\mathbf{E}_{\text{src}}$	$\mathbf{E}_{\text{ant}}$	**E**	**U**	**H**	λ	*err*	*acc*
News								
singular	50	50	400	50	150	10^{-5}	0.086	0.931
plural	50	50	400	50	150	10^{-4}	0.309	0.853
TED								
singular	20	20	160	20	50	10^{-5}	0.131	0.964
plural	50	50	400	50	150	10^{-6}	0.434	0.751

$\mathbf{E}_{\text{src}}$: source embedding size $\mathbf{E}_{\text{ant}}$: antecedent embedding size
E, **U**, **H**: total layer sizes λ: ℓ_2 weight penalty
err: validation error *acc*: accuracy

do not contribute a score, and scoring is left to the translation and language models.

The two neural networks are trained exactly as described in Chapter 8. The network configurations and their intrinsic performance are shown in Table 9.2. They were selected based on validation error after testing a small number of different configurations. To create the training sets for the neural networks, all applicable examples were extracted from the News commentary corpus for the News system and from the TED corpus for the TED system. From these examples, 10 % were held out as a validation set and another 10 % as a test set. The remaining data points, around 7,000 to 8,000 per condition, were combined with examples sampled from the 10^9 corpus to create training sets of about 120,000 examples per text genre and source pronoun.

In the document-level decoder, the 6-gram LM of the baseline system is replaced with a pronoun placeholder LM as described in Section 9.2. Otherwise the feature models are identical. In particular, the bilingual 4-gram LM of the second pass is the same as that of the first pass and does not use placeholders. The same is true of the 9-gram cluster LM, but this makes no difference because the pronouns corresponding to identical placeholders are assigned to the same clusters by the Brown clustering algorithm.

An attempt to optimise the feature weights of the document-level system including the anaphora models failed because document-level MERT against the BLEU score showed no signs of convergence after 25 iterations. We suspect that this failure is due to problems with the sampling procedure that generates the *n*-lists for MERT (see Section 4.7). Instead of tuning the feature weights automatically, we use the same set of weights as for the baseline system and fix the weights of the two anaphora features manually and essentially arbitrarily. The anaphora model weights are set to 0.01 because values of 0.001 and 0.1 result in an unreasonably small or large number of changes in the test set translation. Since we have no reliable automatic performance

metric, we make no attempt at optimising the weights more carefully. While our way of setting parameters based on test set performance without using a separate development set is methodologically objectionable, we consider it very unlikely that this crude method that considers only three different exponentially spaced parameter values and selects the best based on a superficial impression results in a serious unfair advantage for our anaphora model.

To make the anaphora model as effective as possible, it is important for the decoder to be able to change pronoun translations easily. In some cases, a pronoun may be a part of a longer phrase, and it is difficult to alter the entire phrase in a single step without making some accidental changes that cause the modification to be rejected. To give the decoder a chance to make changes in multiple steps, we employ the simulated annealing search algorithm instead of hill climbing. The search is started with a temperature of 1 and follows a slow geometric decay cooling schedule, whereby the temperature is multiplied by 0.99999 after each accepted step. The `crossover` operation (with a weight of 0.2) and the `restore-best` operation (with a weight of 0.1) are used to keep the search from deviating too far from the hill climbing path. The remaining state operations are `change-phrase-translation` (with weight 0.4), `swap-phrases` (with weight 0.2 and swap distance decay 0.5) and `resegment` (with weight 0.1 and phrase size decay 0.1).

For the News corpus, the set of potential antecedents for each occurrence of *it* or *they* is identified with an automatic markable extraction pipeline, and each antecedent candidate is assigned a probability with the neural network exactly as described in Chapter 8. For the TED corpus, we can do the same. Thanks to the existence of the ParCor corpus (Guillou et al., 2014), however, we also have gold-standard pronoun coreference annotations at our disposal. We can therefore run the experiment in a "gold" condition, where we replace the automatically extracted antecedents with the gold-standard information from the ParCor corpus. In this condition, we mark up exactly one antecedent candidate per anaphoric instance of *it* or *they* and assign it a probability of 1. Pronoun occurrences that are marked as non-anaphoric in ParCor are removed. The anaphora models in the "gold" condition are the same as those in the "predicted" condition. In particular, no gold-standard information is used for training the neural networks.

9.3.3 Test Corpora

For the TED system, the test corpus used in our experiments is the *tst2010* test set as distributed in the WIT^3 corpus. It is composed of 11 documents comprising 1,664 segments in total.

In the WMT News test sets, pronouns are distributed very unevenly among the documents. While they are abundant in some documents, others contain very few pronouns or none at all (see Section 6.3 and Table 6.1, p. 95, for some

Table 9.3. *BLEU scores for SMT system with anaphora model*

Corpus	News	TED	
Anaphora resolution	predicted	predicted	gold
Baseline	0.2439	0.3086	
524,288 steps	0.2440	0.3085	0.3079
8,388,608 steps	–	0.3086	0.3080

statistics). To ensure that the phenomena we focus on are sufficiently covered by the test set, we compile a new test set by combining suitable documents from a number of existing test corpora. Our pronoun test corpus is extracted from the *newstest* test sets released for the MT shared tasks at the 2008, 2009, 2010 and 2012 Workshops on Statistical Machine Translation (WMT). The *newstest2011* set is not included because we use it as a development set for feature weight tuning. From these test sets, we extract all documents with at least 5 sentences containing the pronouns *it* or *they* or an uppercase variant of them. The resulting corpus contains 131 documents and 4,954 segments in total. All the News results in this chapter refer to this corpus.

9.3.4 Automatic Evaluation

After the initial decoding run with Moses, we launch Docent with the full set of features including the document-level models. For the TED system, we run Docent for 2^{23} = 8,388,608 steps. For the News system, decoding is much slower because of the larger test set, so we interrupt decoding after 2^{19} = 524,288 steps. After these periods, 360 out of 4,954 segments in the News test set (7.3 %), 122 out of 1,664 segments in the TED experiment with predicted anaphora resolution (7.3 %) and 105 out of 1,664 segments in the TED experiment with gold-standard anaphora resolution (6.3 %) have been modified by the decoder. The slightly lower number of modifications in the "gold" condition may be due to the fact that every anaphor only has a single antecedent candidate in this condition, thus reducing the number of cross-sentence dependencies with respect to the "predicted" condition.

Table 9.3 shows the BLEU scores for these experiments. Clearly, the difference between the baselines and the document-level systems are very small. For the News system and the TED system in the "predicted" condition, the score difference is negligible. For the TED system in the "gold" condition, the score drops by less than 0.1 BLEU points. This change does indicate that the reference translation is matched slightly less closely, but it is too small to permit any conclusions.

The automatic pronoun translation metric introduced in Section 7.3 slightly decreases in both precision and recall for the News texts and for the TED texts

Table 9.4. *Pronoun evaluation scores for SMT system with anaphora model*

	P	R	F
News			
Baseline	0.317	0.343	0.330
predicted	0.313	0.338	0.325
TED			
Baseline	0.451	0.444	0.447
predicted	0.454	0.443	0.449
gold	0.444	0.435	0.440

in the "gold" condition (Table 9.4). For the "predicted" condition of the TED experiment, only recall decreases while precision improves a little, so that the F-score actually increases, but by an entirely negligible amount. These figures do not bode well for our experiments, but we should remember that the automatic metric matches the translation of pronouns against the reference translations without considering the actual anaphoric relations, so its validity is debatable. In sum, the evidence of the automatic scores is neutral or slightly negative, but the negative effects are small and further investigation is warranted nevertheless.

9.4 Manual Pronoun Annotation

To evaluate the performance of our anaphora model in a more focused way, we have developed a manual annotation protocol that allows us to collect gold standard annotations of pronoun choice in machine-translated context. Our annotation scheme generates information that can be used not only for testing how well a given MT system translates pronouns, but also to gain insights about the pronoun evaluation task as such by comparing this evaluation method with a similar method based on reference translations.

Because of the limited time and annotator resources that were available for the manual evaluation, we only evaluated two SMT systems in this way, the News system with predicted anaphoric links and the TED system with gold-standard anaphora annotations. We decided to include one News and one TED system to cover both text types we have systems for. Among the two TED systems, we gave preference to the one with gold-standard annotations even though it differs from the News system in two essential variables because we were unsure if the predicted anaphoric links were sufficiently good and because we conjectured *a priori* that the anaphora model with gold-standard annotations was more likely to have a positive effect on SMT performance.

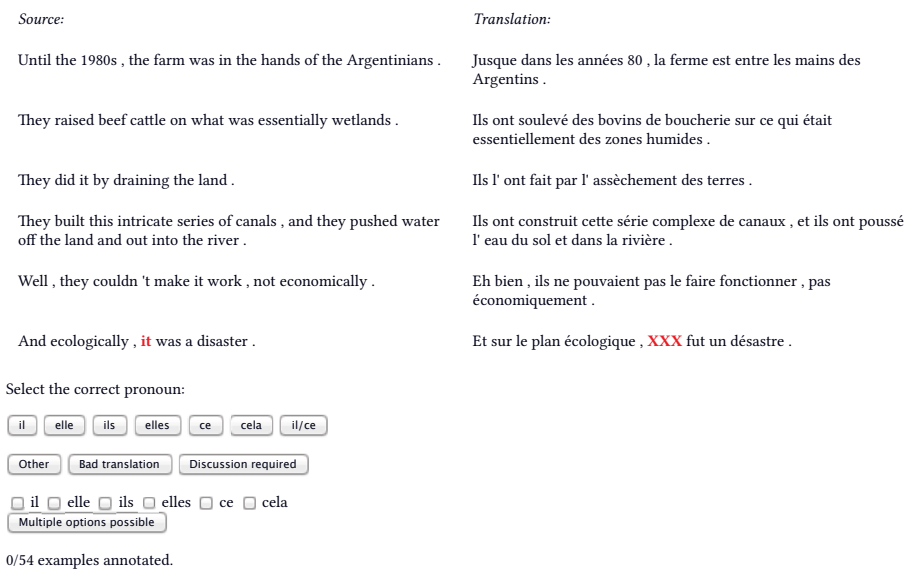

Figure 9.1. Pronoun annotation interface

9.4.1 Annotation Task Description

The main difficulty in evaluating pronoun translations is finding out what the correct translation of a given pronoun is. Usually, MT evaluation assumes that greater overlap between the MT hypothesis and a human-generated reference translation is a sign of better translation quality. When it comes to translating pronouns, this assumption is problematic because a pronoun that matches the reference translation may actually be less correct than another if the context is different. To evaluate pronoun translation correctly, we must therefore find out how the pronoun should be represented in the target language context of the translation, which is exceedingly difficult to do automatically. A human language user, however, can make this decision fairly quickly. The task requires no expert knowledge other than some proficiency in the source and target language.

The annotation work was done through a simple web interface shown in Fig. 9.1. Each example corresponds to one instance of an English pronoun *it* or *they*. The annotators are presented with the sentence containing the pronoun and some preceding context. Up to five sentences of context are included, but fewer if the example is close to the beginning of a document. For all sentences, we also show a translation generated by the MT system to be evaluated. In most sentences, a placeholder is inserted in the MT output of the last sentence containing the pronoun to be annotated. The placeholder replaces any pronoun linked by word alignment to the English target pronoun. As a French pronoun, we consider any word listed with a pronoun

part-of-speech tag (`pro` or any tag starting with `cl`) in the Lefff vocabulary (Sagot et al., 2006). The annotators are then asked to identify the pronoun that should be inserted into the French text instead of the placeholder to create the most fluent translation possible whilst preserving the meaning of the English sentence as much as possible. If no French pronoun is aligned to the English one, no placeholder is inserted, and the annotators are asked to find out if a pronoun corresponding to the one marked up in the English source should be inserted somewhere in the sentence.

The options available to the annotators include six very common French pronouns and three additional categories to mark special cases. The six pronouns are the masculine and feminine singular and plural forms of the subject pronouns, *il*, *elle*, *ils* and *elles*, as well as the pronoun *ce* of the *c'est* 'it is' construction and the frequently used demonstrative pronoun *cela* 'this'. Among these six pronouns, multiple choices are possible if the annotators consider that several equally good completions are available. The three additional categories are named OTHER, representing any pronoun other than the six just mentioned, BAD TRANSLATION, indicating that the machine translation is so bad that it cannot be meaningfully completed with a pronoun, and DISCUSS (called "Discussion required" in the on-line interface) to mark that the annotator is unsure how to handle a specific example. These three categories cannot be combined with each other or any of the pronouns. The annotation interface is designed to permit annotating almost all examples with a single mouse click. Multiple clicks are only necessary if an example should be annotated with more than one pronoun. To allow for one-click annotation in the relatively frequent case where both *il est* and *c'est* are acceptable, we provide a special button named *il/ce*.

Since most of MT output produced by our systems is not perfectly fluent and the additional categories for special cases are very uninformative, we request the annotators to select a pronoun whenever reasonably possible and ignore fluency problems as far as practicable. In particular, they are instructed to disregard any agreement violations that may arise when inserting pronouns for the placeholders. The detailed annotation guidelines are shown in Fig. 9.2. They were shown to the annotators at the beginning of each annotation session and could always be consulted by scrolling down on the web page with the annotation interface.

Annotation work of this type can be carried out fairly quickly at a speed of about one example per minute. Our annotations were created by the author of this thesis, one of his advisors and two colleagues working at the same department.[1] One annotator is a native speaker of French, the others are second-language speakers of French and native speakers of Germanic languages (Swedish or German).

[1]We are indebted to Joakim Nivre, Marie Dubremetz and Mats Dahllöf for their help with the annotations.

For each example, you are presented with up to 5 sentences of English source text and a corresponding French machine translation. In the last sentence, an English pronoun is marked up in red, and (in most cases) the French translation contains a red placeholder for a pronoun. You are asked to select a pronoun that fits in the context.

- Please select the pronoun that should be inserted in the French text instead of the placeholder **XXX** to create the most fluent translation possible while preserving the meaning of the English sentence as much as possible.
- If different, equally grammatical completions are available, select the appropriate checkboxes and click on "Multiple options possible". The button "il/ce" is a special shortcut for cases where these two options are possible.
- Select "Other" if the sentence should be completed with a pronoun not included in the list.
- Select "Bad translation" if there is no way to create a grammatical and faithful translation without making major changes to the surrounding text.
- Select "Discussion required" if you're completely unsure what to do with a particular example.
- Minor disfluencies (e. g., incorrect verb agreement or obviously missing words) can be ignored. For instance, if the placeholder should be replaced with the words *c'est*, just select "ce".
- You should always try to select the pronoun that best agrees with the antecedent in the machine translation, even if the antecedent is translated incorrectly, and even if this forces you to violate the pronoun's agreement with the immediately surrounding words such as verbs, adjectives or participles. So if the antecedent requires a plural form, but the placeholder occurs with a singular verb, you should select the correct plural pronoun and ignore the agreement error.
- If the French translation doesn't contain a placeholder, you should check if a pronoun corresponding to the one marked up in the English source should be inserted somewhere and indicate which if so.
- If the French translation doesn't contain a placeholder, but it already includes the correct pronoun (usually an object pronoun like *le*, *la* or *les*), you should annotate the example as if there had been a placeholder instead of the pronoun (i. e., click on "Other" in the case of an object pronoun).
- Prefer "Bad translation" over "Discussion required" if you're unsure because the translation is dodgy. Reserve "Discussion required" for cases where there is a problem with the guidelines. And don't spend too much thought about the distinction between these two categories, if in doubt, pick the one that came to mind first.

Figure 9.2. Guidelines for the pronoun annotation task

Table 9.5. *Pronoun annotation agreement*

	Exact match				*Overlap*			
Annotator	1	2	3	4	1	2	3	4
1	50	40	37	30	50	44	40	30
2	40	50	33	28	44	50	40	32
3	37	33	50	24	40	40	50	26
4	30	28	24	50	30	32	26	50
Off-diagonal mean	35.7	33.7	31.3	27.3	38.0	38.7	35.3	29.3

9.4.2 Annotation Characteristics

To test the annotation scheme, we collected annotations for a set of 50 examples, of which 24 are taken from the pronoun-enriched News test set and 26 come from the TED test set. The examples were sampled randomly from the two test sets. In total they comprise 32 examples of *it* and 18 examples of *they* (15 *it* and 9 *they* from News data and 17 *it* and 9 *they* from TED data). This set was given to all annotators, so that each example was independently annotated four times. The option to specify multiple pronouns was used very sparingly by the annotators; only 7 out of 200 annotation records make use of this possibility. 12 out of 50 examples (4 News, 8 TED) were labelled with BAD TRANSLATION or DISCUSS by at least one of the annotators. When we inspected these cases after the annotation was completed, we recognised that there was very little difference in the way these two labels were used. The tag DISCUSS almost universally indicates some problem with the translation, and in the following discussion, we make no difference between the two labels. In a very small number of examples, the annotators missed a category NONE to indicate that no pronoun was required. As this was very rare, we decided not to modify the list of categories after creating the initial annotations and use the OTHER category for this purpose instead. In new annotation tasks, however, we recommend adding such a category.

Table 9.5 shows the extent to which the annotators agree. The left part of the table, labelled *Exact match*, contains the number of examples for which the annotations agree exactly. In the right part, labelled *Overlap*, we only require that there should be at least one option that both annotators consider acceptable, regardless of whether one of them also admits other possibilities.

We compute inter-annotator agreement in terms of Krippendorff's α (Krippendorff, 2004) and Scott's π (Scott, 1955) with the software included in the NLTK toolkit (Bird et al., 2009). Over all four annotators, we obtain $\alpha = 0.613$ and $\pi = 0.189$, which suggests significant disagreement. It turns out that a substantial part of the disagreement can be pinned down to a single annotator. If we do not consider the contributions of annotator 4, we reach much better agreement scores of $\alpha = 0.742$ and $\pi = 0.679$. Since it seems accept-

Table 9.6. *Pronoun evaluation contingency table for 80 paired examples*

Baseline	*with anaphora models*							
	News				*TED*			
	−	+	O	B	−	+	O	B
−	11 (10)	3 (1)	0 (0)	1 (1)	11 (10)	0 (0)	1 (1)	0 (0)
+	1 (1)	19 (19)	0 (0)	0 (0)	2 (1)	22 (22)	0 (0)	0 (0)
O	1 (1)	0 (0)	2 (2)	0 (0)	0 (0)	0 (0)	4 (4)	0 (0)
B	0 (0)	0 (0)	1 (1)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)

The figures in parentheses indicate the number of cases with identical pronouns.
−: wrong pronoun +: correct pronoun O: labelled OTHER
B: labelled BAD TRANSLATION or DISCUSS

able to work with three annotators at this level of agreement and we lacked the time for extensive annotator training and guideline revisions, we distribute the examples of the following evaluations in roughly equal shares among annotators 1 to 3, two second-language speakers of French and one native speaker.

9.4.3 Anaphora Model Evaluation

In a first human evaluation round, we annotate a set of 80 example pairs randomly drawn in equal parts from the News commentary and the TED corpus. Each pair consists of a translation created by the baseline SMT system and a translation created by the SMT system with anaphora models, annotated following the guidelines outlined above. Depending on the annotations and the pronoun translation generated by the SMT system, we classify each example into one of four categories. For examples assigned one or more of the labels *il*, *elle*, *ils*, *elles*, *ce* or *cela* by the human annotator, we determine whether the MT system emitted a matching pronoun. Matching is performed case-insensitively. In addition to the six pronouns literally corresponding to the class names, we consider *c'* to be an instance of the class *ce* and *ça* to be an instance of the class *cela*. If the translation of an example with a pronoun label is a match according to these criteria, we classify it as a positive example (+), otherwise as a negative example (−). Examples labelled as OTHER by the human annotators are assigned to class O if the translation generated by the MT system does not correspond to any of the pronoun categories, otherwise to class −, and examples labelled as BAD TRANSLATION or DISCUSS are assigned to class B regardless of the MT output.

Table 9.6 shows contingency tables indicating the classification of the example pairs in the baseline system and in the anaphora-enabled system. The rows of the table correspond to the classes in the baseline output and the columns to the classes in the output of the document-level system. There

Table 9.7. *Contingency table for 88 paired examples with different pronouns*

Baseline	*with anaphora models*							
	News				*TED*			
	−	+	O	B	−	+	O	B
−	9	19	2	0	11	9	2	0
+	11	4	0	0	12	5	0	0
O	0	0	0	0	0	0	1	0
B	1	1	0	0	1	0	0	0

−: wrong pronoun +: correct pronoun O: labelled OTHER
B: labelled BAD TRANSLATION OR DISCUSS

are three factors that can cause an example to migrate from one category to another and end up in an off-diagonal cell of the contingency table. Firstly, the document-level decoder with its anaphora model can alter the translation of a pronoun. Secondly, it can modify the surrounding context or the antecedent translations so that another pronoun becomes appropriate. Such changes may be triggered by the anaphora model or by slight differences between the language models used in the two passes. They could also occur when the local search decoder discovers and corrects search errors made by the baseline decoder. Finally, an example may be assigned to a different category because of inconsistencies in the manual annotation.

In Table 9.6, the anaphora model hardly seems to have any effect. For the purposes of this evaluation, we are primarily interested in the positive and negative examples in the upper left corner of the matrices. Here, only 4 of 34 News examples and 2 of 35 TED examples are categorised differently after running the second-pass decoder. Comparing the pronoun translations in the baseline output with those in the second-pass system output reveals that the pronoun translations are identical in the vast majority of cases. This observation does not enable us to draw any definite conclusions about the behaviour of the system because the correctness of the pronoun translation depends, in addition to the pronoun itself, on the context and the antecedent translations. However, it does raise suspicion that the translations of the baseline and of the anaphora-enabled system may be equivalent in many cases. To evaluate our anaphora model, we need to know whether its effect is positive or negative in those cases where there actually is an effect.

As an approximation to the examples influenced by the anaphora model, we consider the subset of examples where the final translation after document-level decoding has a translation of the pronoun which is different from that of the baseline. This is the case for 151 out of 1,457 News examples and 63 out of 566 TED examples, considering only examples where the source pronoun is aligned to a pronoun in the target language in both the baseline and

the document-level system. Pronoun comparisons are performed in a case-sensitive manner and only exact literal matches are considered equal because anything else but a literal, case-sensitive match indicates a motivated choice by the SMT system. From this subset, we consider a random sample of 47 News examples and 41 TED examples. The results are reported in Table 9.7.

In 60 of 88 example pairs (68.2 %), at least one of the two translations produced by the baseline and the document-level system matches the preference of the annotators. Additionally, some items in the O class may be correct as well. However, the number of items assigned to the classes O and B is small, so we concentrate our discussion on the positive and negative pronoun classes (– and +). In the News text genre, the number of examples migrating from – to + (19 items) is distinctly larger than the number of examples moving in the opposite direction (11 items). However, the difference is not large enough to be significant at a 90 % confidence level in Liddell's test (Liddell, 1983). Surprisingly enough, in the TED data, where gold-standard coreference annotations are used, the anaphora model seems to cause some damage, with 12 items going from + to – and only 9 from – to +. Needless to say, this difference is far from being statistically significant.

Considering the small sample size and the absence of statistical significance, we cannot rule out the possibility that the effects we observe are due to random variations. Even so, the relatively large positive effect in the News experiment attracts attention, and so does the unexpected negative outcome of the TED experiment. In our opinion, both results deserve further investigation. Most importantly, the manual evaluation should be continued with larger samples than the time constraints for this thesis have permitted us to examine. This will allow us to test if the effects persist and become significant or if they must be dismissed as random. Additionally, assuming the effect observed in the News experiment is confirmed, a similar evaluation should be conducted for the "predicted" condition of the TED experiment to find out if it bears more resemblance to the "predicted" condition of the News experiment or to the "gold" condition of the TED experiment. Both hypotheses are possible.

On the one hand, the BLEU scores in Table 9.3 suggest a greater similarity between the two "predicted" conditions than between the two conditions of the TED experiment. This is a highly dubious indication because the score differences are quite small and we have strong reasons to distrust BLEU as a measure of pronoun translation accuracy. However, if the TED experiment should prove more successful with predicted anaphora resolution than with gold-standard annotations, this would be a very interesting result. The mismatch between training and testing conditions could be a possible explanation for such a finding. At training time, the distribution over antecedent candidates encoded by the network's **V** layer will generally have a fairly large entropy because of the great uncertainty of the anaphora resolution process. It is not impossible that the unexpected use of a very sharp distribution con-

centrating all its probability mass on a single item at testing time has unintended effects on the operation of the network, even if the distribution is known to be correct.

On the other hand, even if the positive effect in the News experiment subsists at larger sample sizes, it may be more difficult to achieve comparable performance for the text genre encountered in the TED talks. In the intrinsic evaluations of Chapter 8, we found that the pronouns in the TED data are easier to predict than those in the News data. However, this may well be due to the fact that there is less entropy in the prior distribution of the pronouns, as evidenced by the better accuracy of the majority class baseline (Table 8.3, p. 120). Despite their superior overall performance, it is not clear that the TED networks actually perform better when predicting more difficult edge cases. However, the prior distribution of the pronouns should already be matched well by the language model of the baseline SMT system, so there may be less room for improvement in the TED experiment.

9.4.4 Agreement with Reference Translation

With the manual annotations created to evaluate the anaphora model and the human reference translations of the test sets, we have two very different and mutually independent types of gold-standard information on the translation of pronouns in our test corpora. The reference translations indicate how a text, including the pronouns it contains, can be translated in a correct manner. We assume that the human translators producing the translations have a good understanding of the source text and of the target language norms and create high-quality output even if there is no one-to-one correspondence between source language and target language elements, and even if target language conventions dictate pronoun usage patterns that are not strictly consistent with source language usage. However, the correctness of the reference translation can only be guaranteed for the translation as a whole. If only some bits and pieces of a candidate translation tally with the reference translation while other parts diverge, we cannot be sure that the total result will be acceptable.

The manual annotations, by contrast, are specifically created for a particular candidate translation generated by an MT system. Even if that candidate translation as a whole is inferior to the reference translation, the pronoun translation suggested by the manual annotations is more reliable than the one suggested by the reference translation because it is consistent with the context of the machine translation. From a theoretical point of view, the translation of a pronoun found in the reference text cannot be a valid solution in the MT context other than by chance. However, in practice, reference translations are routinely used to calculate automatic quality scores for all parts of MT output, including the translations of input pronouns. It is therefore pertinent to examine to what extent the translations of pronouns in a reference

translation corpus are useful to evaluate candidate translations produced by an MT system.

To compare reference translations with manual annotations, we first create a set of pseudo-annotations based on the references. We generate word alignments between the reference translations and the input texts by concatenating them with the parallel training corpus and running the same word alignment procedure that we use for training the SMT system. Then we construct an annotation record for every occurrence of the pronouns *it* and *they* in the input, setting the label in accordance with the target language element aligned to the input pronoun. Again, *c'* is counted as an instance of of *ce* and *ça* is counted as an instance of *cela*. Comparisons are performed case-insensitively. No pseudo-annotation record is created if the input pronoun is not aligned to a target language word in the reference translation.

The first thing to notice with these automatically generated pseudo-annotations is that many pronoun occurrences are not covered by them. Of the 1,547 examples of *it* or *they* extracted from the News reference translation, 517 (33.4 %) are not aligned to a pronoun in the target language.[2] In the TED data, 245 of 735 examples (33.3 %) are not aligned to pronouns. A superficial manual inspection of the data reveals that the word alignment is usually correct in these cases. Most often, these are genuine examples of translations where pronoun usage differs between the source and the target language. This does not necessarily mean that it would be impossible to translate the input in a way that preserves the pronoun usage of the source language while respecting the target language norms, but it makes it impossible to evaluate these examples with the information contained in the reference translations and greatly reduces the usefulness of any evaluation scheme that relies on reference translations for pronoun evaluation. This observation applies to the automatic pronoun evaluation metric of Section 7.3 as well as to the pseudo-annotations considered here.

Because of the great number of examples for which simple pronoun correspondences cannot be extracted from the reference translation, the effective sample size of the pronoun evaluation is reduced and it becomes more difficult to appraise the significance of an effect. Moreover, very likely the subset of source pronouns that are not aligned to a target pronoun is not randomly drawn from the total set of source pronouns, so ignoring it will bias the evaluation. Conversely, it could be argued that this subset will incorporate all the examples for which a pronoun translation is linguistically impossible, so it contains valuable information about how to translate those cases. This is true, but since these will be examples where the reference translation deviates substantially from the wording of the input, the word alignment will be

[2]The total number of examples extracted varies slightly across translations because we skip examples when a source pronoun is aligned to more than one target pronoun. This occurs most frequently when *it is* is rendered as *il y a* 'there is' and *it* is aligned to the pronouns *il* and *y*.

Table 9.8. *Contingency table for 88 paired examples with different pronouns, evaluated with pseudo-annotations*

Baseline	*with anaphora models*							
	News				*TED*			
	–	+	O	B	–	+	O	B
–	11	10	1	–	11	6	1	–
+	8	0	0	–	13	1	0	–
O	0	0	0	–	0	0	0	–
B	–	–	–	–	–	–	–	–

–: wrong pronoun +: correct pronoun O: labelled OTHER
B: labelled BAD TRANSLATION or DISCUSS

unreliable, and the information will be far from trivial to exploit. Also, current SMT models are highly unlikely to generate good output for these cases, whereas they may well occasionally produce acceptable output for examples where a direct pronominal translation is possible even if the creator of the reference translation decided against using it.

Table 9.8 shows the results obtained by evaluating the 88 example pairs used in Table 9.7 with the pseudo-annotations generated from human reference translations. The category B is not used in this table because the pseudo-annotations never carry the labels BAD TRANSLATION or DISCUSS. It turns out that the pseudo-annotations are less likely to classify an example as correct (+) than the human-made annotations specific to the MT system, which must be considered as the gold standard in this comparison. The effect applies both to the News and to the TED system. In both cases, the document-level system is affected more strongly than the baseline. This may be an effect of chance, but it would have led to an overly negative evaluation of the anaphora model in this case.

In Table 9.9, the 176 manual annotations collected for the same 88 example pairs are pitted against the corresponding pseudo-annotations. Since the manual annotations come in pairs, each pseudo-annotation occurs twice in this table in combination with two different manual annotations. If we consider only the first two rows, where there is either a clear match or a clear mismatch with the manual annotation, the pseudo-annotation matches the manual annotation in only 43 of 90 News cases (47.8 %). For 33 examples (36.7 %), there is no pseudo-annotation, and in 14 cases (15.6 %), the pseudo-annotations flatly contradict the judgements of the human annotators. The figures for the TED data are considerably better with 50 matches in 77 examples (64.9 %), 18 missing annotations (23.4 %) and 9 contradictions (11.7 %), but even in the TED corpus, pseudo-annotations are either incorrect or missing for more than one third of the examples.

Table 9.9. *Contingency table for manual annotations versus pseudo-annotations*

manual annotations	*pseudo-annotations*							
	News				*TED*			
	−	+	O	∅	−	+	O	∅
−	29	4	0	18	32	3	0	11
+	10	14	0	15	6	18	0	7
O	0	0	1	1	3	0	1	0
B	2	0	0	0	1	0	0	0

−: wrong pronoun +: correct pronoun O: labelled OTHER
B: labelled BAD TRANSLATION or DISCUSS ∅: no annotation

The results suggest that the pseudo-annotations, and very probably also other reference-oriented measures such as our pronoun evaluation metric of Section 7.3 and BLEU, misrepresent the correctness of anaphora translations and will not do justice to improvements achieved by specific anaphora handling components. The severity of the problem is corpus-dependent, but it is clearly present in both of the corpora we have examined. This finding confirms the theoretically motivated hypothesis that reference-oriented measures are insufficient to guide the development of systems modelling complex target-side dependencies.

9.5 Conclusion

In the experiments in this chapter, we have tested the pronoun prediction model developed in the previous chapter in practical SMT systems for two different text genres. While the model has no effect on the automatic evaluation scores, manual evaluation of the News experiment with predicted anaphora resolution reveals a mildly positive result in that the number of improvements exceeds the number of regressions by a small margin. This result, however, is modest and uncertain, and it is not borne out by the parallel experiment on TED data with gold-standard anaphora resolution. Nevertheless, the SMT implementation of the anaphora model and its subsequent evaluation have unfolded a number of interesting insights.

First of all, the experiments afford a new confirmation that the decoding algorithm developed in the first part of this thesis is viable for practical use. After examining the output of the document-level decoder, we have no reason to suppose that the limited success of the anaphora model is due to the fact that the decoder fails to improve the model scores. Rather, the shortcomings we observe can be pinned down convincingly to difficulties of the task and inadequacies of the feature models.

Another important insight is the recognition that it is mistaken to assume a direct correspondence of pronouns across languages. This fact has not yet been internalised sufficiently by the SMT community. Early work on pronouns in SMT, including our own, naïvely assumed that pronouns were anaphoric as a rule and that anaphoric pronouns, barring rare exceptions, could be mapped directly onto corresponding target language pronouns (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010). This is not what we find in corpus data. Even though we recognised this problem before developing our pronoun prediction model (Chapter 6), it turns out that the capacity of our model to cope with it is still insufficient and that more sophisticated modelling will be required for an adequate solution.

The results of the manual evaluation of our anaphora model, while inconclusive, are intriguing. In the News experiment with predicted anaphoric links, we observe a small improvement over the baseline. The improvement is not statistically significant, but it is strong enough to nurture hope that it will survive and prove significant when larger samples are studied. By contrast, and quite contrary to what we originally expected, we find no improvement in an experiment with TED data and gold-standard coreference annotations. We have advanced some speculations as to why this might be the case, but only more empirical work can show to what extent they are true.

Finally, the comparison of manual and automatic evaluations for our anaphora model has uncovered deficiencies in the automatic evaluation procedures that were already known in theory, but whose actual impact had not been demonstrated empirically. Based on the results presented in this chapter, we can state with some confidence that BLEU and other reference-oriented evaluation measures are insufficient tools for the development of models of pronominal anaphora and similar phenomena involving complex target-side dependencies. Currently, we cannot suggest a better automatic evaluation score for this purpose, but the manual evaluation protocol described above permits the collection of targeted and more reliable annotations at a relatively low cost.

10. Conclusions

In this thesis, we address discourse-level aspects of translation in phrase-based SMT from different points of view. Throughout our work, we have been confronted with both technical and linguistic challenges. The technical challenges are related to the independence assumptions made by existing SMT solutions, and correspondingly, our first research goal has been to *develop frameworks, procedures and algorithms that are not encumbered by the standard assumptions of sentence-level independence.* As a response to this challenge, we have developed and explored a framework for document-level decoding and released the Docent decoder (Hardmeier et al., 2013a). The linguistic challenges, on the other hand, are related to our second research goal, to *investigate what discourse-level linguistic phenomena can be useful for SMT, and how to model them in an SMT system.* We have studied different types of discourse-level information, but our principal effort is dedicated to the problem of pronoun translation. We investigate the behaviour of pronouns under translation, present a neural network model to predict the French translations of English pronouns and integrate this model into a phrase-based SMT system. In this chapter, we recapitulate the findings of our thesis, discuss the insights gained and contributions made and highlight some issues that should be addressed in future work.

10.1 Document-Level SMT

In the first part of the thesis, we study the technical problems that we encounter when integrating document-level features into SMT decoding. We show how the widely used stack decoding algorithm exploits locality assumptions to speed up decoding with a dynamic programming technique called recombination, which makes it unsuitable for use with features that have long-range dependencies. Decoding with such features requires special techniques to overcome the independence assumptions of the decoding algorithm. We discuss three methods that enable us to combine document-level features with sentence-level decoding algorithms, by decoding in two passes, by propagating information between sentences during a single decoding pass or by running a second-pass search with a different algorithm over a subset of the search space represented by the n-best output of a stack decoder. All of these methods have been used in the literature. They trade off modelling

constraints, search space, ease of implementation and efficiency against each other in different ways.

The core contribution of the first part of this thesis is the development of a new local search decoder for phrase-based SMT at the document level. It embodies a new approach to decoding with document-level models which makes trade-offs that are very different from those of the existing methods. The assumption of sentence independence is radically removed and the modelling constraints it causes, such as the dependency directionality constraint in the information propagation approaches, are lifted. The search space accessible to the local search decoder is equal, at least in principle, to the full search space of phrase-based SMT.

As regards ease of implementation, the decoding algorithm is geared towards complex document-level models. Simple models with local dependencies such as an n-gram language model can be considerably more complicated to implement in the document-level local search framework than in a DP stack decoder because the stack decoder constructs its output in an order that is particularly well suited to n-gram-style dependencies and all the required information is readily available. The local search decoder, by contrast, gives the programmer complete freedom to define the dependencies of the model, and it is about as difficult to define a model with remote dependencies across sentence boundaries as one with local dependencies only.

The most important trade-off made by the document-level decoder is that of efficiency. By exploiting the locality of the models with dynamic programming, the traditional stack decoder manages to explore a comparatively large part of the search space with relatively little effort, even though it still has to resort to pruning to ensure polynomial runtime. The local search decoder does not have this advantage and potentially spends much more time covering an equivalent part of the search space. It is important to remember, however, that the stack decoder's efficiency advantage is tightly coupled to the locality of the models. It only exists in a condition in which the local search decoder is not designed to be used. As soon as the locality constraints on the models are softened and long-range dependencies are admitted in the models, the DP technique in the stack decoder becomes less effective, and its head start begins to vanish. If the dependencies are left completely unconstrained, DP is no longer applicable and the stack decoder will not necessarily be more efficient than the local search decoder any more.

Fusing the efficiency of stack decoding with the versatility of document-level local search, we show that the local search decoder can be initialised with a search state obtained from a stack decoder. In this setup, the DP search of the first pass solves a relaxed version of the decoding problem from which the constraints involving long-range dependencies have been omitted. While there is no theoretical guarantee that the state found by DP search with the relaxed models is a good starting point for the document-level search, it is reasonable to assume that it is generally better than a random point in the

search space, especially if the overall model of the document-level search pass is relatively similar to that of the DP search pass. We test this decoding setup with different discourse-level models, including a semantic space language model, a collection of readability models and a pronominal anaphora model. We have not evaluated these experiments specifically for decoding performance, but clearly the decoder is capable of improving the model score and even of overfitting to peculiarities of the models in all cases, and we find no indications of fundamental problems with the search method in any of the experiments. We conclude that local search with DP initialisation is a viable solution for experimenting with discourse-level models in phrase-based SMT.

One of the principal benefits of having a decoder that admits unlimited document-level dependencies, and our main motivation for creating and releasing this piece of software, is that it enables researchers to experiment freely with discourse-level models without imposing technical restrictions on the space of imaginable models from the beginning. The availability of a document-level decoding framework should make it possible to test ideas that would otherwise be abandoned in an early stage because the expected cost of implementation is considered too high in relation to the probability of success. Once a particular method has been demonstrated to work and is ready to be incorporated into a production system, other techniques than local search may prove more effective depending on the nature of the model.

In the work presented in this thesis, we show that the local search method works for phrase-based SMT decoding, but we do not explore its parameters very thoroughly, accentuating instead the development of discourse-level feature models. Now that a number of models have been developed, there are many aspects of the search process that merit closer attention. The acceptance criterion of the local search algorithm lends itself as a starting point. Hill climbing reliably directs the search towards higher-scoring regions of the search space, but theoretical considerations suggest that it may fail to find optimal solutions because it requires a score improvement at each individual search step. However, some improvements may only be achievable if the decoder is permitted to make one or more intermediate steps to states with lower scores first, e. g., to split up a phrase pair into smaller pieces that can be manipulated independently.

To enable the decoder in a principled way to explore search paths in which the model scores do not increase monotonically, we can employ the stochastic Metropolis-Hastings acceptance criterion and perform simulated annealing instead of hill climbing. In initial experiments with simulated annealing not reported in this thesis, there was evidence of significant problems. Even when started in a relatively high-scoring state, the decoder would quickly abandon promising regions of the search space and wander off towards very bad states without finding its way back to any acceptable solutions within a reasonable period of time.

We surmise that these search problems are connected with the set of search operations we use, and particularly with the fact that the combination of the proposal distribution and the acceptance criterion of our decoder does not satisfy the elementary theoretical conditions guaranteeing convergence of the simulated annealing procedure. However, by adding operations that tie the decoder to the hill climbing path and limit the duration of excursions to lower-scoring regions of the search space, the effectiveness of simulated annealing search can be greatly increased despite the persistence of the theoretical difficulties. In future work, the design and selection of search operations for both hill climbing and simulated annealing and the interaction between the proposal distribution and the search algorithm in simulated annealing should be investigated more thoroughly and with greater focus on theoretical convergence results.

Another problem that urgently needs more attention is feature weight tuning. Stymne et al. (2013a) present an adaptation of the MERT algorithm (Och, 2003) to document-level decoding, also described in Section 4.7 of this thesis, and show that it achieves useful results under certain circumstances. However, when we try to apply the same method to our system with pronominal anaphora models in Chapter 9, MERT completely fails to converge. We conjecture that this failure is due to poor sampling parameters in the generation of n-lists, but owing to time constraints we could not study the problem more closely.

Instead of using MERT, feature weight optimisation could be performed with the PRO method (Hopkins and May, 2011) that estimates the weights as parameters of a linear classifier trained to separate good states encountered by the decoder from bad ones. Stymne et al. (2013a) do test PRO with the sampling method they also use for MERT, but training data for PRO could potentially also be collected by making the decoder search directly for a state with optimal BLEU score or, preferably, some other measure of translation quality more sensitive to discourse-level aspects of translation quality. This option is currently being explored in ongoing work at Uppsala University.

In sum, there are still a number of issues related to document decoding that require further study, but already now, the decoding method we present has proved to be an enabling factor for a number of experiments with discourse-level models including the work on anaphora in the second part of the thesis, which demonstrates its usefulness at least as a research tool.

10.2 Pronominal Anaphora in SMT

In the second part of this thesis, we turn to the issue of pronominal anaphora. We start by examining the translations of pronouns in German–English MT output and verify that pronoun translation is, in fact, a problem for SMT. We find that the adequacy of pronoun translations varies greatly across different

types of pronouns and, as a function of the prevalence of certain pronoun types in the individual documents, across documents. The overall accuracy is on the order of 60 % and considerably lower for some pronoun types affected by morphological syncretism with other more frequent forms in the source language such as feminine singular pronouns in German. Depending on the contents of the documents translated, such pronouns may be rare, but pervasive mistranslation of particular types of pronouns is vexatious for the reader and may even create an appearance of disrespect, especially if there is a noticeable gender bias in the way pronouns are translated (Gendered Innovations, 2014). We therefore conclude that pronoun translation is a problem with some practical impact in current state-of-the-art SMT.

Having established this fact, we discuss a number of complications that arise when modelling pronominal anaphora in an SMT system. The pronoun translation task is complex and requires doing inference over information collected from a number of sources and resulting from a variety of components, each of which suffers from uncertainty and is liable to add a certain amount of noise to the system. Since each of the individual components involves highly complex reasoning, it easily happens that the accumulated noise drowns all useful information in the system.

After a brief description of an early approach to pronoun modelling in SMT and its evaluation, we introduce a neural network classifier that models cross-lingual pronoun prediction as a task in its own right, independently of an MT system. In terms of raw accuracy, the neural network improves a bit over a simple maximum entropy classifier. However, the improvement is not very large, presumably because the distribution of the pronouns in the data is heavily skewed so that it is relatively easy to attain high accuracy just by predicting the most frequent classes more frequently; for one of the two text genres tested, the accuracy of the maximum entropy classifier is only marginally higher than that of a trivial majority choice baseline. Still, the neural network has considerable advantages over the baseline because it delivers acceptable precision and recall for all output classes, whereas the baseline only performs well for the more frequent target language pronouns. In particular, it greatly improves the prediction performance for the French feminine plural pronoun *elles*. We use *elles* as an indicator of progress, because to predict this pronoun correctly, the classifier must exploit information from the antecedents of the pronouns and cannot rely on unconditional frequency distributions and the immediate context of the pronouns alone.

An important feature of our neural network classifier is its capability to model the links between anaphoric pronouns and their antecedents as latent variables, eliminating the need for an external coreference resolution system trained on manually annotated data. Instead, we extend the network with a small number of extra layers to model the probability of anaphoric links given a set of features prepared with the feature extraction machinery of the existing anaphora resolver. We then train these layers jointly with the pronoun

prediction layers by backpropagating the error gradients all the way from the pronoun prediction network into the anaphoric link scoring component, using unannotated parallel text as the only supervision. The fact that this approach works just as well as using the predictions of the external coreference resolution system reveals that parallel bitexts contain valuable information about pronominal coreference that had never been exploited in SMT prior to our work, and only to a small extent in coreference resolution research.

We conclude our experimental work by incorporating the pronoun prediction neural network as a feature model into the document-level local search decoder. In doing so, we tie together all the major contributions of this thesis. We test the resulting system on two text types, news data and TED talks. For the TED talks, we have access to a test set with manually created annotations of pronominal coreference, which gives us the opportunity to examine the performance of this system both with the latent anaphora resolution of the neural network and with the gold-standard anaphoric links in the manually annotated data set.

In terms of automatic quality measures, the anaphora model has very little effect on the performance of the SMT systems. The BLEU score remains all but unchanged for all systems, and our own automatic pronoun evaluation metric is inconclusive as well. If anything, it is surprising that the TED system with gold-standard anaphora resolution fares worse than the corresponding system with predicted anaphoric links, but the score differences are far too small to draw conclusions with any degree of confidence.

Since we are well aware that the existing automatic evaluation measures are inherently unreliable when it comes to studying pronoun translation, we conduct a simple and rapid manual evaluation of two of our systems with a small number of annotators, which provides us with information on the most adequate translation of pronouns in the actual context of MT output. The evaluation yields very interesting, if somewhat inconclusive results. We observe an improvement in pronoun translation for the News corpus with predicted anaphoric links, but not for the TED corpus with gold-standard annotations. While neither of the results is statistically significant, the outcome for the News corpus is strong enough to inspire hope that significance might be attained if a larger sample were examined. The negative result in the TED experiment tallies with the marginally negative result of the automatic evaluation and raises the intriguing question whether the difference, if indeed there is a difference in substance, is due to the features of the two text genres tested or to the fact that the neural network trained for unsupervised anaphora resolution is confused by the presence of gold-standard annotation.

At present, all of this is mere speculation because the observed effects are very modest and chance is a factor to be reckoned with considering the small samples we have examined. Even so, we believe the results are interesting enough to warrant further investigation in future work. Furthermore, although the work presented in this thesis has not led to a breakthrough in

terms of translation quality, it has shed some light on the difficulties involved in translating anaphoric pronouns with an SMT system.

First of all, we must recognise that pronoun translation is more difficult than it seems, and more difficult than has been acknowledged by most SMT researchers who have even made an effort to solve it. The complications discussed in Chapter 6 are confirmed anew by the experimental results of Chapter 8 and Chapter 9, and despite being aware of many of the challenges when designing these experiments, we have not been able to avert all the problems they cause.

The existing research on pronouns in SMT has largely concentrated on the problems of resolving pronominal anaphora, identifying the translation of the antecedent and injecting the information gained through anaphora resolution into an SMT system. These are essential steps without which we cannot hope to solve the pronoun translation problem. Aside from relying on the effects of chance and the skewness of pronoun distributions, there is no way around the fact that correctly generating a pronoun like the French *elles* requires information about the translation of its antecedent, and obtaining this information is difficult and has justly been the object of some research efforts. However, what has been underestimated so far is that pronoun translation is a challenging discourse problem even if we leave aside the problem of coreference resolution completely, and that it is qualitatively different from translating content words.

Different languages have different conventions of pronoun use, and the translation of pronouns is subject to arbitrary effects of linguistic conventions to a much greater extent than the translation of content words. Consider, by way of example, the case of company names, which is relatively frequent in news texts. In English, as in other languages, companies are frequently introduced with their name:

(10.1) a. A perfidious embezzler. This is how the French banking giant Société Générale, the owner of the local Komerční banka (Commerce Bank), labels its ex-employee Jerome Kerviel.

b. Un fraudeur dissimulateur. Ainsi désigne son ancien employé le géant français la Société générale, propriétaire de la banque tchèque Komerční banka. (*news-test2008*)

In English, it is then common to refer back to the company name using the pronoun *it*. In French, by contrast, it is often more idiomatic to refer to the company name with a full noun phrase first, although it is not strictly impossible to use a pronoun directly:

(10.2) a. On his account *it* has lost almost five billion Euro.

b. *La banque* a perdu à cause de lui près de cinq milliards d'euros. (*news-test2008*)

The following example exhibits two completely different complications. On the one hand, it uses a highlighting idiom that is specific to the English language and must be rendered with other means in French. On the other hand, an English subordinate clause is mapped into a construction involving a present participle which does not require an explicit subject pronoun.

(10.3) a. But the thing about tryptamines is *they* cannot be taken orally because *they*'re denatured by an enzyme found naturally in the human gut called monoamine oxidase.

b. Par contre les tryptamines ne peuvent pas être consommées par voie orale étant dénaturé[e]s par une enzyme se trouvant de façon naturelle dans l'intestin de l'homme : la monoamine-oxydase.[1]

Note that both instances of the English word *they* are regular anaphoric pronouns with a clearly defined antecedent, yet neither of these pronouns occurs in the French reference translation. Moreover, translating a subordinate clause with a finite verb and a pronominal subject into a participle or gerund without overt subject is frequently possible in different language pairs, also when English is the target language.

There is evidence suggesting that cases like these are far more common in bilingual corpus data than one might believe. In addition to the anecdotic examples we have presented here and in other places in this thesis, the overwhelming predominance of the OTHER class in the training data of the neural networks presented in Chapter 8 (Table 8.2, p. 119) and the great number of English pronouns not aligned to French pronouns in the pseudo-annotations of Section 9.4.4 indicate that it is fairly common for pronouns not to be rendered literally in translation, even though those figures may incorporate other special cases such as incorrect word alignments as well.

Now it could be argued that the translators creating these reference translations take excessive liberties with the input text and that they should be instructed to translate more literally at least when producing reference translations for SMT research. However, this argument is fallacious. By requesting more literal translations, we would force the translators to translate "verbum e verbo", in the manner recognised to be inadequate already by the church father Jerome in the fourth century (Jerome, 1996). A consistently more literal rendering would amount to word glossing, not translation, and it would have a strongly negative impact at least on the idiomaticity, if not on the fluency of the target language text. Moreover, creating artificially literal reference translations for SMT use could have a lasting negative impact on the progress of MT research because evaluating against these references would favour the overly literal translation style of existing models while penalising more sophisticated systems that may be developed in the future.

[1] This example is taken from the *dev2010* test set of the WIT^3 corpus (Cettolo et al., 2012).

Rather than artificially simplifying the reference data, the only sustainable, if challenging, way to cope with these difficulties is to analyse the relevant phenomena and attempt to model them adequately. We expect that future approaches to pronoun translation in SMT will require extensive corpus analysis to study how pronouns of a given source language are rendered in a given target language and create a classification of these instances. While it may not be possible to explain all cases satisfactorily with the means currently at our disposal, much would be gained if we could identify with some confidence which cases are amenable to handling with our existing models to prevent the system from introducing spurious errors in the remaining cases.

10.3 Final Remarks

The recent work on discourse in SMT, and the difficulties we and others have experienced when trying to improve MT with discourse models, reveal some basic weaknesses of the SMT approach. Most current approaches to SMT are founded on word alignments in the spirit of Brown et al. (1990). These word alignments have no clear theoretical status. They are defined in terms of statistical models whose parameters are estimated based on cooccurrence statistics extracted from a training corpus, and they mirror a concept of translational equivalence that we have termed observational equivalence to distinguish it from the higher-level notion of dynamic equivalence and its counterpart, formal equivalence, of which it could be considered a special case.

Observational equivalence is strongly surface-oriented, and SMT has traditionally eschewed all abstract representations of meaning, mapping tokens of the input directly into tokens of the output. This has worked well, demonstrating that much linguistic information is indeed accessible with surface-level processing. However, one problem of this approach is that the SMT system often does not know exactly what it is doing. For instance, based on observational evidence from the training corpus, an SMT system might translate an active sentence in the input with a passive sentence in the output, or a personal construction in the source language with an impersonal construction in the target language.

In English–French translation, this happens not infrequently, e. g., when the phrase *it requires* is translated with the impersonal *il faut* ‘it is necessary’, *it* being aligned to *il*. In this example, the English *it* is anaphoric, but the French *il* is pleonastic. This translation may be perfectly adequate and idiomatic, as the training data suggests, but the problem is that the SMT system has no control over what it is doing. Just copying bits and pieces of texts that it encountered at training time, it does not know that a personal pronoun is being mapped into an impersonal one in the example above, or that the subject and object functions are exchanged in a sentence when it goes from active to passive.

It is difficult to envisage consistently correct translation of discourse phenomena such as pronominal anaphora or generating the correct distribution of definite and indefinite noun phrases if the MT system is not allowed to construct any abstract representation of the entities occurring in a text. In some way, future SMT systems will have to make inferences about more abstract entities than surface words to create adequate translations. This could be done with the help of a capacity for symbolic reasoning over some form of abstract semantic representation (e. g., Banarescu et al., 2013), but it is not clear that symbolic representations are, in fact, the most suitable approach. Quite possibly, abstract information about texts could be represented in the form of one or more hidden layers in a neural network or a similar latent-variable representation (e. g., along the lines of Kalchbrenner and Blunsom, 2013). Creating such a mechanism, and making it interface with the existing surface-level processing facilities, is going to be a major research effort and is unlikely to lead to improvements in BLEU score in the short term.

We began this thesis by drawing attention to a discrepancy between translation studies and SMT research, pointing out how the two fields are concerned with challenges at entirely different levels of abstraction. We showed that the observational equivalence aimed at in SMT corresponds to a fairly dated view of translation that misses out not only the cultural turn of the late 20th century in translation studies and the shift of viewpoint towards seeing translation as a procedural phenomenon in a cultural and social context, but even earlier developments of the concept of equivalence such as the functional notion of dynamic equivalence advocated by Nida and Taber (1969). It is now time to reflect what this thesis has contributed to promote a more up-to-date concept of translation in SMT research.

First of all, the limitations of our efforts must be clearly acknowledged. None of our contributions will bring about a paradigm shift from a view of SMT focusing on translational equivalence to a more process-oriented view of translation, nor have we even attempted to do so. While it is important to keep in mind that the underlying assumptions of current approaches to SMT fall short of the insights prevalent in modern translation research, we believe that it is appropriate, given the current state of the art, that SMT should rest on a concept of equivalence and that matters related to the intentionality of the source text and its translation and to their social or cultural context should be regarded as external to the SMT translation process itself. If anything, we should wish that the nature of this equivalence relation as well as the concept of domain, which encodes many of those external factors, were put on a firmer theoretical basis. This, however, has not been the subject of this thesis.

What we have attempted to do is to free phrase-based SMT from the narrow-minded focus on *n*-gram context and sentence independence and create a framework in which modelling at a larger scale is possible without being impeded by technical constraints from the very beginning. We consider that

this is an enabling factor to promote research on the translation of linguistic phenomena on the text level, but also on aspects of SMT emanating from a broader view of translational equivalence. In the applications treated in this thesis, both can be found. Pronominal anaphora, which we devoted most effort to, is an example of an elementary linguistic phenomenon that requires discourse-level processing for correct translation even if no more than mere formal equivalence is called for. By contrast, the readability experiments briefly discussed in Section 5.2 represent an effort that transcends even the limits of dynamic equivalence by conferring on the translation an intention not found in the source text and retargeting the text to a new audience.

In sum, notwithstanding the practical contributions we have made, the foremost importance of this thesis is theoretical rather than practical. By highlighting the fundamental limitations of one of the prevalent approaches to SMT, by studying their impact on practical translations and by creating a new framework relaxing the most stringent restrictions and demonstrating its applicability to unresolved issues in MT, we hope to stimulate SMT research with a greater propensity for creating explanatory models of complex textual relations.

Bibliography

Aarts, Emile H. L., Korst, Jan H. M. and van Laarhoven, Peter J. M. (1997). Simulated annealing. In: Emile H. L. Aarts and Jan Karel Lenstra (eds.), *Local Search in Combinatorial Optimization*, Wiley-Interscience series in discrete mathematics and optimization, Chichester: Wiley, 91–120.

Alexandrescu, Andrei and Kirchhoff, Katrin (2009). Graph-based learning for statistical machine translation. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder (Colorado, USA), 119–127.

Artstein, Ron and Poesio, Massimo (2008). Inter-coder agreement for computational linguistics. *Computational linguistics*, 34 (4):555–596.

Arun, Abhishek, Dyer, Chris, Haddow, Barry, Blunsom, Phil, Lopez, Adam and Koehn, Philipp (2009). Monte Carlo inference and maximization for phrase-based translation. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, Boulder (Colorado, USA), 102–110.

Arun, Abhishek, Haddow, Barry, Koehn, Philipp, Lopez, Adam, Dyer, Chris and Blunsom, Phil (2010). Monte Carlo techniques for phrase-based translation. *Machine translation*, 24 (2):103–121.

Baker, Mona (2011). *In other words. A coursebook on translation.* London: Routledge. Second edition.

Banarescu, Laura, Bonial, Claire, Cai, Shu, Georgescu, Madalina, Griffitt, Kira, Hermjakob, Ulf, Knight, Kevin, Koehn, Philipp, Palmer, Martha and Schneider, Nathan (2013). Abstract meaning representation for sembanking. In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, Sofia (Bulgaria), 178–186.

Banchs, Rafael E. and Costa-jussà, Marta R. (2011). A semantic feature for statistical machine translation. In: *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Portland (Oregon, USA), 126–134.

Banerjee, Satanjeev and Lavie, Alon (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor (Michigan, USA), 65–72.

Bassnett, Susan (2011). The translator as cross-cultural mediator. In: Kirsten Malmkjær and Kevin Windle (eds.), *The Oxford Handbook of Translation Studies*, Oxford: Oxford University Press, 94–107.

Becher, Viktor (2011a). *Explicitation and implicitation in translation. A corpus-based study of English–German and German–English translations of business texts.* Ph. D. thesis, Universität Hamburg.

Becher, Viktor (2011b). When and why do translators add connectives? A corpus-based study. *Target: International Journal on Translation Studies*, 23 (1):26–47.

Beigman Klebanov, Beata, Diermeier, Daniel and Beigman, Eyal (2008). Lexical cohesion analysis of political speech. *Political Analysis*, 16 (4):447–463.

Beigman Klebanov, Beata and Flor, Michael (2013). Associative texture is lost in translation. In: *Proceedings of the Workshop on Discourse in Machine Translation*, Sofia (Bulgaria), 27–32.

Bellegarda, Jerome R. (2000). Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE*, 88 (8):1279–1296.

Ben, Guosheng, Xiong, Deyi, Teng, Zhiyang, Lü, Yajuan and Liu, Qun (2013). Bilingual lexical cohesion trigger model for document-level machine translation. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia (Bulgaria), 382–386.

Bengio, Yoshua, Ducharme, Réjean, Vincent, Pascal and Janvin, Christian (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Berger, Adam L., Della Pietra, Stephen A. and Della Pietra, Vincent J. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22 (1):39–72.

Bergsma, Shane and Yarowsky, David (2011). NADA: A robust system for non-referential pronoun detection. In: *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium*, Faro (Portugal), *Lecture Notes in Computer Science*, volume 7099, 12–23.

Bird, Steven, Loper, Edward and Klein, Ewan (2009). *Natural Language Processing with Python*. Beijing: O'Reilly.

Björnsson, Carl-Hugo (1968). *Läsbarhet*. Stockholm: Liber.

Brown, Peter F., Cocke, John, Della Pietra, Stephen A., Della Pietra, Vincent J., Jelinek, Frederick, Lafferty, John D., Mercer, Robert L. and Roossin, Paul S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16 (2):79–85.

Brown, Peter F., Della Pietra, Stephen A., Della Pietra, Vincent J. and Mercer, Robert L. (1993). The mathematics of statistical machine translation. *Computational linguistics*, 19 (2):263–311.

Brown, Peter F., deSouza, Peter V., Mercer, Robert L., Della Pietra, Vincent J. and Lai, Jenifer C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18 (4):467–479.

Buch-Kromann, Matthias, Korzen, Iørn and Høeg Müller, Henrik (2009). Uncovering the 'lost' structure of translations with parallel treebanks. *Copenhagen Studies in Language*, 38:199–224.

Bungum, Lars and Gambäck, Björn (2011). A survey of domain adaptation in machine translation: Towards a refinement of domain space. In: *Proceedings of the India-Norway Workshop on Web Concepts and Technologies*, Trondheim (Norway).

Bussmann, Hadumond (1996). *Routledge dictionary of language and linguistics*. Routledge Reference, London: Routledge.

Callison-Burch, Chris, Koehn, Philipp, Monz, Christof, Peterson, Kay, Przybocki, Mark and Zaidan, Omar (2010). Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala (Sweden), 17–53.

Callison-Burch, Chris, Koehn, Philipp, Monz, Christof, Post, Matt, Soricut, Radu and Specia, Lucia (2012). Findings of the 2012 Workshop on Statistical Machine Translation. In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montréal (Canada), 10–51.

Callison-Burch, Chris, Koehn, Philipp, Monz, Christof and Schroeder, Josh (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens (Greece), 1–28.

Callison-Burch, Chris, Koehn, Philipp, Monz, Christof and Zaidan, Omar (2011). Findings of the 2011 Workshop on Statistical Machine Translation. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh (Scotland, UK), 22–64.

Callison-Burch, Chris, Osborne, Miles and Koehn, Philipp (2006). Re-evaluating the role of BLEU in machine translation research. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento (Italy).

Carpuat, Marine (2009). One translation per discourse. In: *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, Boulder (Colorado, USA), 19–27.

Carpuat, Marine and Simard, Michel (2012). The trouble with SMT consistency. In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montréal (Canada), 442–449.

Carpuat, Marine and Wu, Dekai (2007). Improving statistical machine translation using word sense disambiguation. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague (Czech Republic), 61–72.

Cartoni, Bruno, Zufferey, Sandrine and Meyer, Thomas (2013). Annotating the meaning of discourse connectives by looking at their translation: The translation spotting technique. *Dialogue and Discourse*, 4 (2):65–86.

Cartoni, Bruno, Zufferey, Sandrine, Meyer, Thomas and Popescu-Belis, Andrei (2011). How comparable are parallel corpora? Measuring the distribution of general vocabulary and connectives. In: *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, Portland (Oregon, USA), 78–86.

Cettolo, Mauro, Girardi, Christian and Federico, Marcello (2012). WIT^3: Web inventory of transcribed and translated talks. In: *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento (Italy), 261–268.

Chall, Jeanne S. (1958). *Readability: An appraisal of research and application.* Columbus (Ohio): Bureau of Educational Research.

Chen, Stanley F. and Goodman, Joshua (1998). An empirical study of smoothing techniques for language modeling. Technical Report, Computer Science Group, Harvard University, Cambridge (Mass.).

Chiang, David (2007). Hierarchical phrase-based translation. *Computational linguistics*, 33 (2):201–228.

Chiang, David (2012). Hope and fear for discriminative training of statistical translation models. *Journal of Machine Learning Research*, 13:1159–1187.

Cho, Eunah, Ha, Thanh-Le, Mediani, Mohammed, Niehues, Jan, Herrmann, Teresa, Slawik, Isabel and Waibel, Alex (2013). The Karlsruhe Institute of Technology translation systems for the WMT 2013. In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia (Bulgaria), 104–108.

Coccaro, Noah and Jurafsky, Daniel (1998). Towards better integration of semantic predictors in statistical language modeling. In: *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney (Australia).

Collins, Michael (1999). *Head-Driven Statistical Models for Natural Language Parsing*. Ph. D. thesis, University of Pennsylvania.

Collins, Michael and Duffy, Nigel (2002). Convolution kernels for natural language. In: Thomas G. Dietterich, Suzanna Becker and Zoubin Ghahramani (eds.), *Advances in Neural Information Processing Systems 14*, Cambridge (Mass.): MIT Press, 625–632.

Collobert, Ronan, Weston, Jason, Bottou, Léon, Karlen, Michael, Kavukcuoglu, Koray and Kuksa, Pavel (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2461–2505.

Deléger, Louise, Merkel, Magnus and Zweigenbaum, Pierre (2006). Enriching medical terminologies: An approach based on aligned corpora. In: Arie Hasman, Reinhold Haux, Johan van der Lei, Etienne De Clercq and Francis H. Roger France (eds.), *Ubiquity: Technologies for Better Health in Aging Societies. Proceedings of MIE2006, the 20th International Congress of the European Federation for Medical Informatics*, Maastricht (Netherlands), 747–752.

Denkowski, Michael and Lavie, Alon (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh (Scotland, UK), 85–91.

Doddington, George (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: *Proceedings of the Second International Conference on Human Language Technology Research*, San Diego (California, USA), 138–145.

Eidelman, Vladimir, Boyd-Graber, Jordan and Resnik, Philip (2012). Topic models for dynamic translation model adaptation. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Jeju Island (Korea), 115–119.

Eisner, Jason and Tromble, Roy W. (2006). Local search with very large-scale neighborhoods for optimal permutations in machine translation. In: *Proceedings of the HLT-NAACL Workshop on Computationally Hard Problems and Joint Inference in Speech and Language Processing*, New York City (New York, USA), 57–75.

Evans, Richard (2001). Applying machine learning toward an automatic classification of *it*. *Literary and Linguistic Computing*, 16 (1):45–57.

Federico, Marcello, Bertoldi, Nicola and Cettolo, Mauro (2008). IRSTLM: An open source toolkit for handling large scale language models. In: *Interspeech 2008*, Brisbane (Australia), 1618–1621.

Federmann, Christian, Eisele, Andreas, Chen, Yu, Hunsicker, Sabine, Xu, Jia and Uszkoreit, Hans (2010). Further experiments with shallow hybrid MT systems. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala (Sweden), 77–81.

Fellbaum, Christiane (1998). *WordNet: An electronic lexical database*. Cambridge (Mass.): MIT Press.

Finkel, Jenny Rose, Grenager, Trond and Manning, Christopher (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor (Michigan, USA), 363–370.

Foltz, Peter W., Kintsch, Walter and Landauer, Thomas K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25 (2/3):285–307.

Foster, George, Isabelle, Pierre and Kuhn, Roland (2010). Translating structured documents. In: *Proceedings of AMTA 2010: the Ninth Conference of the Association for Machine Translation in the Americas*, Denver (Colorado, USA).

Gale, William A., Church, Kenneth W. and Yarowsky, David (1992). One sense per discourse. In: *Proceedings of Speech and Natural Language*, Harriman (New York, USA), 233–237.

Galley, Michel and McKeown, Kathleen (2003). Improving word sense disambiguation in lexical chaining. In: *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, San Francisco (California, USA), 1486–1488.

Gendered Innovations (2014). Machine translation: Analyzing gender. `http://genderedinnovations.stanford.edu/case-studies/nlp.html` (5 May 2014).

Germann, Ulrich (2003). Greedy decoding for statistical machine translation in almost linear time. In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton (Canada).

Germann, Ulrich, Jahr, Michael, Knight, Kevin, Marcu, Daniel and Yamada, Kenji (2001). Fast decoding and optimal decoding for machine translation. In: *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse (France), 228–235.

Germann, Ulrich, Jahr, Michael, Knight, Kevin, Marcu, Daniel and Yamada, Kenji (2004). Fast and optimal decoding for machine translation. *Artificial Intelligence*, 154 (1–2):127–143.

Giménez, Jesús, Màrquez, Lluís, Comelles, Elisabet, Castellón, Irene and Arranz, Victoria (2010). Document-level automatic MT evaluation based on discourse representations. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala (Sweden), 333–338.

Gong, Zhengxian, Zhang, Min, Tan, Chew-lim and Zhou, Guodong (2012a). Classifier-based tense model for SMT. In: *Proceedings of COLING 2012: Posters*, Mumbai (India), 411–420.

Gong, Zhengxian, Zhang, Min, Tan, Chew Lim and Zhou, Guodong (2012b). N-gram-based tense models for statistical machine translation. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island (Korea), 276–285.

Gong, Zhengxian, Zhang, Min and Zhou, Guodong (2011a). Cache-based document-level statistical machine translation. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh (Scotland, UK), 909–919.

Gong, Zhengxian, Zhang, Yu and Zhou, Guodong (2010). Statistical machine translation based on LDA. In: *Proceedings of the 4th International Universal Communication Symposium*, Beijing (China), 286–290.

Gong, Zhengxian, Zhou, Guodong and Li, Liangyou (2011b). Improve SMT with source-side "topic-document" distributions. In: *Proceedings of the 13th Machine Translation Summit*, Xiamen (China), 496–501.

Grishman, Ralph and Sundheim, Beth (1996). Message understanding conference – 6: A brief history. In: *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, Copenhagen (Denmark), 466–471.

Gruber, Amit, Weiss, Yair and Rosen-zvi, Michal (2007). Hidden topic Markov models. In: Marina Meila and Xiaotong Shen (eds.), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-07)*, San Juan (Puerto Rico, USA), volume 2, 163–170.

Guillou, Liane (2011). *Improving Pronoun Translation for Statistical Machine Translation (SMT).* Master's thesis, University of Edinburgh, School of Informatics.

Guillou, Liane (2012). Improving pronoun translation for statistical machine translation. In: *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon (France), 1–10.

Guillou, Liane (2013). Analysing lexical consistency in translation. In: *Proceedings of the Workshop on Discourse in Machine Translation*, Sofia (Bulgaria), 10–18.

Guillou, Liane, Hardmeier, Christian, Smith, Aaron, Tiedemann, Jörg and Webber, Bonnie (2014). ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In: *Proceedings of the Tenth Language Resources and Evaluation Conference (LREC'14)*, Reykjavík (Iceland).

Gupta, Kamakhyn, Sadiq, Mohamed and Sridhar V (2008). Measuring lexical cohesion in a document. In: *Seventh Mexican International Conference on Artificial Intelligence*, Tuxtla Gutiérrez (Mexico), 48–52.

Guzmán, Francisco, Joty, Shafiq, Màrquez, Lluís and Nakov, Preslav (2014). Using discourse structure improves machine translation evaluation. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore (Maryland, USA).

Hajič, Jan, Panevová, Jarmila, Hajičová, Eva, Panevová, Jarmila, Sgall, Petr, Pajas, Petr, Štěpánek, Jan, Havelka, Jiří and Mikulová, Marie (2006). *Prague Dependency Treebank 2.0*. Philadelphia: Linguistic Data Consortium. LDC2006T01.

Hajlaoui, Najeh and Popescu-Belis, Andrei (2012). Translating English discourse connectives into Arabic: a corpus-based analysis and an evaluation metric. In: *AMTA-2012: Fourth workshop on computational approaches to Arabic script-based languages*, San Diego (California, USA), 1–8.

Hajlaoui, Najeh and Popescu-Belis, Andrei (2013). Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In: Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, Berlin: Springer, *Lecture Notes in Computer Science*, volume 7817, 236–247.

Halliday, M. A. K. and Hasan, Ruqaiya (1976). *Cohesion in English*. English Language Series, London: Longman.

Harabagiu, Sanda M. and Maiorano, Steven J. (2000). Multilingual coreference resolution. In: *Proceedings of the Sixth Conference on Applied Natural Language Processing*, Seattle (Washington, USA), 142–149.

Hardmeier, Christian (2012). Discourse in statistical machine translation: A survey and a case study. *Discours*, 11.

Hardmeier, Christian, Bisazza, Arianna and Federico, Marcello (2010). FBK at WMT 2010: Word lattices for morphological reduction and chunk-based reordering. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala (Sweden), 88–92.

Hardmeier, Christian and Federico, Marcello (2010). Modelling pronominal anaphora in statistical machine translation. In: *Proceedings of the Seventh International Workshop on Spoken Language Translation (IWSLT)*, Paris (France), 283–289.

Hardmeier, Christian, Nivre, Joakim and Tiedemann, Jörg (2012). Document-wide decoding for phrase-based statistical machine translation. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island (Korea), 1179–1190.

Hardmeier, Christian, Stymne, Sara, Tiedemann, Jörg and Nivre, Joakim (2013a). Docent: A document-level decoder for phrase-based statistical machine translation. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Sofia (Bulgaria), 193–198.

Hardmeier, Christian, Stymne, Sara, Tiedemann, Jörg, Smith, Aaron and Nivre, Joakim (2014). Anaphora models and reordering for phrase-based SMT. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore (Maryland, USA).

Hardmeier, Christian, Tiedemann, Jörg and Nivre, Joakim (2013b). Latent anaphora resolution for cross-lingual pronoun prediction. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle (Washington, USA), 380–391.

Hardmeier, Christian, Tiedemann, Jörg, Saers, Markus, Federico, Marcello and Prashant, Mathur (2011). The Uppsala-FBK systems at WMT 2011. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh (Scotland, UK), 372–378.

Hasler, Eva, Blunsom, Phil, Koehn, Philipp and Haddow, Barry (2014). Dynamic topic adaptation for phrase-based MT. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg (Sweden), 328–337.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57 (1):97–109.

Hatim, Basil and Mason, Ian (1990). *Discourse and the Translator*. Language in Social Life Series, London: Longman.

Heafield, Kenneth (2011). KenLM: faster and smaller language model queries. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh (Scotland, UK), 187–197.

Hopkins, Mark and May, Jonathan (2011). Tuning as ranking. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh (Scotland, UK), 1352–1362.

Huang, Yan (2004). Anaphora and the pragmatics-syntax interface. In: Laurence R. Horn and Gregory Ward (eds.), *The Handbook of Pragmatics*, Malden (Mass.): Blackwell, 288–314.

Hultman, Tor G. and Westman, Margareta (1977). *Gymnasistsvenska*. Lund: LiberLäromedel.

Jelinek, Frederick (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64 (4):532–557.

Jellinghaus, Michael, Poulis, Alexandros and Kolovratník, David (2010). Exodus – Exploring SMT for EU institutions. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala (Sweden), 110–114.

Jerome (1979). Letter LVII: To Pammachius on the best method of translating. In: *St. Jerome: Letters and Select Works*, Grand Rapids: Eerdmans, *A Select Library of Nicene and Post-Nicene Fathers of the Christian Church, Second Series*, volume VI, 112–119.

Jerome (1996). Epistola LVII: Ad Pammachium. De optimo genere interpretandi. In: *Sancti Eusebii Hieronymi Stridonensis presbyteri epistolae secundum ordinem temporum ad amussim digestae et in quatuor classes distributae*, Alexandria: Chadwyck-Healey, *Patrologia Latina Database*, volume 22.

Johnson, Howard, Martin, Joel, Foster, George and Kuhn, Roland (2007). Improving translation quality by discarding most of the phrasetable. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague (Czech Republic), 967–975.

Joty, Shafiq, Guzmán, Francisco, Màrquez, Lluís and Nakov, Preslav (2014). DiscoTK: Using discourse structure for machine translation evaluation. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore (Maryland, USA).

Jurgens, David and Stevens, Keith (2010). The S-Space package: An open source package for word space models. In: *Proceedings of the ACL 2010 System Demonstrations*, Uppsala (Sweden), 30–35.

Kalchbrenner, Nal and Blunsom, Phil (2013). Recurrent continuous translation models. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle (Washington, USA), 1700–1709.

Kamp, Hans and Reyle, Uwe (1993). *From Discourse to Logic: An Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer.

Kim, Woosung and Khudanpur, Sanjeev (2004). Cross-lingual latent semantic analysis for language modeling. In: *Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, Montréal (Canada), volume 1, 257–260.

Kirkpatrick, S., Gelatt Jr., C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220 (4598):671–680.

Knight, Kevin and Chander, Ishwar (1994). Automated postediting of documents. In: *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI)*, Seattle (Washington, USA), 779–784.

Koehn, Philipp (2005). Europarl: A corpus for statistical machine translation. In: *Proceedings of MT Summit X*, Phuket (Thailand), 79–86.

Koehn, Philipp (2010). *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Koehn, Philipp and Hoang, Hieu (2007). Factored translation models. In: *Conference on Empirical Methods in Natural Language Processing*, Prague (Czech Republic), 868–876.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra et al. (2007). Moses: Open source toolkit for Statistical Machine Translation. In: *Annual Meeting of the Association for Computational Linguistics: Demonstration session*, Prague (Czech Republic), 177–180.

Koehn, Philipp, Och, Franz Josef and Marcu, Daniel (2003). Statistical phrase-based translation. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Edmonton (Canada), 48–54.

Koller, Werner (1972). *Grundprobleme der Übersetzungstheorie, unter besonderer Berücksichtigung schwedisch-deutscher Übersetzungsfälle*, *Acta Universitatis Stockholmiensis. Stockholmer germanistische Forschungen*, volume 9. Bern: Francke.

Krippendorff, Klaus (2004). Measuring the reliability of qualitative text analysis data. *Quality and Quantity*, 38 (6):787–800.

Lambert, Patrik, Gispert, Adriá, Banchs, Rafael and Mariño, José B. (2005). Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39 (4):267–285.

Langlais, Philippe, Patry, Alexandre and Gotti, Fabrizio (2007). A greedy decoder for phrase-based statistical machine translation. In: *TMI-2007: Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, Skövde (Sweden), 104–113.

Langlais, Philippe, Patry, Alexandre and Gotti, Fabrizio (2008). Recherche locale pour la traduction statistique par segments. In: *Actes de la 15e Conférence sur le Traitement Automatique des Langues Naturelles*, Avignon (France), 119–128.

Le, Hai-Son, Allauzen, Alexandre and Yvon, François (2012). Continuous space translation models with neural networks. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal (Canada), 39–48.

Le Nagard, Ronan and Koehn, Philipp (2010). Aiding pronoun translation with co-reference resolution. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala (Sweden), 252–261.

Leal, Alice (2012). Equivalence. In: Yves Gambier and Luc van Doorslaer (eds.), *Handbook of Translation Studies*, Amsterdam: Benjamins, volume 3, 39–46.

Lee, Heeyoung, Peirsman, Yves, Chang, Angel, Chambers, Nathanael, Surdeanu, Mihai and Jurafsky, Dan (2011). Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, Portland (Oregon, USA), 28–34.

Lefevere, André and Bassnett, Susan (1995). Introduction: Proust's grandmother and the thousand and one nights: The 'cultural turn' in translation studies. In: Susan Bassnett and André Lefevere (eds.), *Translation, History and Culture*, London: Cassell, 1–14.

Liddell, F. D. K. (1983). Simplified exact analysis of case-referent studies: Matched pairs; dichotomous exposure. *Journal of Epidemiology and Community Health*, 37 (1):82–84.

Louis, Annie and Webber, Bonnie (2014). Structured and unstructured cache models for SMT domain adaptation. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg (Sweden), 155–163.

Ma, Yanjun, He, Yifan, Way, Andy and van Genabith, Josef (2011). Consistent translation using discriminative learning – A translation memory-inspired approach. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland (Oregon, USA), 1239–1248.

Mann, William and Thompson, Sandra (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8 (3):243–281.

Marcu, Daniel, Carlson, Lynn and Watanabe, Maki (2000). The automatic translation of discourse structures. In: *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, Seattle (Washington, USA), 9–17.

Mariño, José, Banches, Rafael E., Crego, Josep M., de Gispert, Adrià, Lambert, Patrik, Fonollosa, José A. R. and Costa-jussà, Marta R. (2006). N-gram-based machine translation. *Computational linguistics*, 32 (4):527–549.

McEnery, Anthony, Tanaka, Izumi and Botley, Simon (1997). Corpus annotation and reference resolution. In: *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, Madrid (Spain), 67–74.

Mediani, Mohammed, Cho, Eunah, Niehues, Jan, Herrmann, Teresa and Waibel, Alex (2011). The KIT English–French translation systems for IWSLT 2011. In: *Proceedings of the Eighth International Workshop on Spoken Language Translation (IWSLT)*, San Francisco (California, USA), 73–78.

Merkel, Magnus (1999). *Understanding and enhancing translation by parallel text processing*. Ph. D. thesis, Linköping University.

Metropolis, Nicholas, Rosenbluth, Arianna W., Rosenbluth, Marshall N., Teller, Augusta H. and Teller, Edward (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21 (6):1987–1092.

Meyer, Thomas (2011). Disambiguating temporal-contrastive connectives for machine translation. In: *Proceedings of the ACL 2011 Student Session*, Portland (Oregon, USA), 46–51.

Meyer, Thomas, Grisot, Cristina and Popescu-Belis, Andrei (2013). Detecting narrativity to improve English to French translation of simple past verbs. In: *Proceedings of the Workshop on Discourse in Machine Translation*, Sofia (Bulgaria), 33–42.

Meyer, Thomas and Poláková, Lucie (2013). Machine translation with many manually labeled discourse connectives. In: *Proceedings of the Workshop on Discourse in Machine Translation*, Sofia (Bulgaria), 43–50.

Meyer, Thomas and Popescu-Belis, Andrei (2012). Using sense-labeled discourse connectives for statistical machine translation. In: *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, Avignon (France), 129–138.

Meyer, Thomas, Popescu-Belis, Andrei, Hajlaoui, Najeh and Gesmundo, Andrea (2012). Machine translation of labeled discourse connectives. In: *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego (California, USA).

Meyer, Thomas, Popescu-Belis, Andrei, Zufferey, Sandrine and Cartoni, Bruno (2011a). Multilingual annotation and disambiguation of discourse connectives for machine translation. In: *Proceedings of the SIGDIAL 2011 Conference*, Portland (Oregon, USA), 194–203.

Meyer, Thomas, Roze, Charlotte, Cartoni, Bruno, Danlos, Laurence, Zufferey, Sandrine and Popescu-Belis, Andrei (2011b). Disambiguating discourse connectives using parallel corpora. In: *Proceedings of Corpus Linguistics*, Birmingham (England, UK).

Meyer, Thomas and Webber, Bonnie (2013). Implicitation of discourse connectives in (machine) translation. In: *Proceedings of the Workshop on Discourse in Machine Translation*, Sofia (Bulgaria), 19–26.

Minnen, Guido, Carroll, John and Pearce, Darren (2001). Applied morphological processing of English. *Natural Language Engineering*, 7 (3):207–223.

Mitchell, Alexis, Strassel, Stephanie, Przybocki, Mark, Davis, J. K., Doddington, George, Grishman, Ralph, Meyers, Adam, Brunstein, Ada, Ferro, Lisa and Sundheim, Beth (2003). *ACE-2 Version 1.0*. Philadelphia: Linguistic Data Consortium. LDC2003T11.

Mitkov, Ruslan and Barbu, Catalina (2003). Using bilingual corpora to improve pronoun resolution. *Languages in Contrast*, 4 (2):201–211.

Mühlenbock, Katarina and Johansson Kokkinakis, Sofie (2009). LIX 68 revisited: An extended readability measure. In: *Proceedings of Corpus Linguistics*, Liverpool (England, UK).

Ng, Vincent (2010). Supervised noun phrase coreference research: The first fifteen years. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala (Sweden), 1396–1411.

Nida, Eugene A. and Taber, Charles R. (1969). *The theory and practice of translation, Helps for translators*, volume 8. Leiden: Brill.

Niehues, Jan, Herrmann, Teresa, Vogel, Stephan and Waibel, Alex (2011). Wider context by using bilingual language models in machine translation. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh (Scotland, UK), 198–206.

Novák, Michal (2011). Utilization of anaphora in machine translation. In: *Week of Doctoral Students 2011 Proceedings of Contributed Papers, Part I*, Prague (Czech Republic), 155–160.

Novák, Michal, Nedoluzhko, Anna and Žabokrtský, Zdeněk (2013a). Translation of "it" in a deep syntax framework. In: *Proceedings of the Workshop on Discourse in Machine Translation*, Sofia (Bulgaria), 51–59.

Novák, Michal, Žabokrtský, Zdeněk and Nedoluzhko, Anna (2013b). Two case studies on translating pronouns in a deep syntax framework. In: *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya (Japan), 1037–1041.

Och, Franz Josef (2003). Minimum error rate training in Statistical Machine Translation. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo (Japan), 160–167.

Och, Franz Josef and Ney, Hermann (2002). Discriminative training and maximum entropy models for Statistical Machine Translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia (Pennsylvania, USA), 295–302.

Och, Franz Josef and Ney, Hermann (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29 (1):19–51.

Och, Franz Josef and Ney, Hermann (2004). The alignment template approach to statistical machine translation. *Computational linguistics*, 30 (4):417–449.

Och, Franz Josef, Ueffing, Nicola and Ney, Hermann (2001). An efficient A* search algorithm for Statistical Machine Translation. In: *Proceedings of the Data-Driven Machine Translation Workshop at the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, Toulouse (France), 55–62.

Paice, C. D. and Husk, G. D. (1987). Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun "it". *Computer Speech and Language*, 2 (2):109–132.

Papineni, Kishore, Roukos, Salim, Ward, Todd and Zhu, Wei-Jing (2002). BLEU: A method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia (Pennsylvania, USA), 311–318.

Petrov, Slav, Barrett, Leon, Thibaux, Romain and Klein, Dan (2006). Learning accurate, compact, and interpretable tree annotation. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney (Australia), 433–440.

Petrov, Slav and Klein, Dan (2007). Improved inference for unlexicalized parsing. In: *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester (New York, USA), 404–411.

Popescu-Belis, Andrei, Cartoni, Bruno, Gesmundo, Andrea, Henderson, James, Grisot, Cristina, Merlo, Paola, Meyer, Thomas, Moeschler, Jacques and Zufferey, Sandrine (2012a). Improving MT coherence through text-level processing of input texts: The COMTIS project. In: *Tralogy, Session 6 – Translation and Natural Language Processing / Traduction et traitement automatique des langues (TAL)*, Paris (France).

Popescu-Belis, Andrei, Meyer, Thomas, Liyanapathirana, Jeevanthi, Cartoni, Bruno and Zufferey, Sandrine (2012b). Discourse-level annotation over Europarl for machine translation: Connectives and pronouns. In: *Proceedings of the Eighth Language Resources and Evaluation Conference (LREC'12)*, Istanbul (Turkey).

Postolache, Oana, Cristea, Dan and Orăsan, Constantin (2006). Transferring coreference chains through word alignment. In: *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC-2006)*, Genoa (Italy), 889–892.

Pradhan, Sameer, Ramshaw, Lance, Marcus, Mitchell, Palmer, Martha, Weischedel, Ralph and Xue, Nianwen (2011). CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, Portland (Oregon, USA), 1–27.

Rahman, Altaf and Ng, Vincent (2012). Translation-based projection for multilingual coreference resolution. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal (Canada), 720–730.

Ruiz, Nick and Federico, Marcello (2011). Topic adaptation for lecture translation through bilingual latent semantic models. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh (Scotland, UK), 294–302.

Rumelhart, David E., Hinton, Geoffrey E. and Williams, Ronald J. (1986). Learning representations by back-propagating errors. *Nature*, 323 (6088):533–536.

Russo, Lorenza, Loáiciga, Sharid and Gulati, Asheesh (2012a). Improving machine translation of null subjects in Italian and Spanish. In: *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon (France), 81–89.

Russo, Lorenza, Loáiciga, Sharid and Gulati, Asheesh (2012b). Italian and Spanish null subjects. A case study evaluation in an MT perspective. In: *Proceedings of the Eighth Language Resources and Evaluation Conference (LREC'12)*, Istanbul (Turkey), 1779–1784.

Russo, Lorenza, Scherrer, Yves, Goldman, Jean-Philippe, Loáiciga, Sharid, Nerima, Luka and Wehrli, Éric (2011). Étude inter-langues de la distribution et

des ambiguïtés syntaxiques des pronoms. In: Mathieu Lafourcade and Violaine Prince (eds.), *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles*, Montpellier (France), volume 2, 279–284.

Sagot, Benoît, Clément, Lionel, Villemonte de La Clergerie, Éric and Boullier, Pierre (2006). The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. In: *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC-2006)*, Genoa (Italy), 1348–1351.

Sanders, T. and Pander Maat, H. (2006). Cohesion and coherence: Linguistic approaches. In: *Encyclopedia of language and linguistics*, Elsevier, volume 2, 591–595.

Savoy, Jacques (1999). A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science*, 50 (10):944–952.

Scherrer, Yves, Russo, Lorenza, Goldman, Jean-Philippe, Loáiciga, Sharid, Nerima, Luka and Wehrli, Éric (2011). La traduction automatique des pronoms. Problèmes et perspectives. In: Mathieu Lafourcade and Violaine Prince (eds.), *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles*, Montpellier (France), volume 2.

Schmid, Helmut and Laws, Florian (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In: *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester (England, UK), 777–784.

Schwenk, Holger (2007). Continuous space language models. *Computer Speech and Language*, 21 (3):492–518.

Scott, William A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19 (3):321–325.

Shannon, Claude E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27 (3):379–423.

Snell-Hornby, Mary (1995). Linguistic transcoding or cultural transfer? A critique of translation theory in Germany. In: Susan Bassnett and André Lefevere (eds.), *Translation, History and Culture*, London: Cassell, 79–86.

Snell-Hornby, Mary (2010). The turns of translation studies. In: *Handbook of Translation Studies*, Amsterdam: John Benjamins, volume 1, 366–370.

Snover, Matthew, Dorr, Bonnie, Schwartz, Richard, Micciulla, Linnea and Makhoul, John (2006). A study of translation edit rate with targeted human annotation. In: *AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, "Visions for the Future of Machine Translation"*, Cambridge (Massachussetts, USA), 223–231.

Soon, Wee Meng, Ng, Hwee Tou and Lim, Daniel Chung Yong (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27 (4):521–544.

de Souza, José and Orăsan, Constantin (2011). Can projected chains in parallel corpora help coreference resolution? In: Iris Hendrickx, Sobha Lalitha Devi, António Branco and Ruslan Mitkov (eds.), *Anaphora Processing and Applications*, Berlin: Springer, *Lecture Notes in Computer Science*, volume 7099, 59–69.

Stolcke, Andreas (2002). SRILM: An extensible language modeling toolkit. In: *Proceedings of the International Conference on Spoken Language Processing*, Denver (Colorado, USA).

Stolcke, Andreas, Zheng, Jing, Wang, Wen and Abrash, Victor (2011). SRILM at sixteen: Update and outlook. In: *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa (Hawaii, USA).

Stuckardt, Roland (2007). Applying backpropagation networks to anaphor resolution. In: António Branco (ed.), *Anaphora: Analysis, Algorithms and Applications. 6th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007*, Lagos (Portugal), *Lecture Notes in Artificial Intelligence*, volume 4410, 107–124.

Stymne, Sara, Hardmeier, Christian, Tiedemann, Jörg and Nivre, Joakim (2013a). Feature weight optimization for discourse-level SMT. In: *Proceedings of the Workshop on Discourse in Machine Translation*, Sofia (Bulgaria), 60–69.

Stymne, Sara, Hardmeier, Christian, Tiedemann, Jörg and Nivre, Joakim (2013b). Tunable distortion limits and corpus cleaning for SMT. In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia (Bulgaria), 225–231.

Stymne, Sara, Tiedemann, Jörg, Hardmeier, Christian and Nivre, Joakim (2013c). Statistical machine translation with readability constraints. In: Stephan Oepen, Kristin Hagen and Janne Bondi Johannesse (eds.), *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, Oslo (Norway), 375–386.

Taira, Hirotoshi, Sudoh, Katsuhito and Nagata, Masaaki (2012). Zero pronoun resolution can improve the quality of J-E translation. In: *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Jeju Island (Korea), 111–118.

Tam, Yik-Cheung, Lane, Ian and Schultz, Tanja (2007). Bilingual LSA-based adaptation for statistical machine translation. *Machine Translation*, 21 (4):187–207.

Tiedemann, Jörg (2010a). Context adaptation in statistical machine translation using models with exponentially decaying cache. In: *Proceedings of the ACL 2010 Workshop on Domain Adaptation for Natural Language Processing (DANLP)*, Uppsala (Sweden), 8–15.

Tiedemann, Jörg (2010b). To cache or not to cache? Experiments with adaptive models in statistical machine translation. In: *Proceedings of the ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala (Sweden), 189–194.

Trask, R. L. (1993). *A dictionary of grammatical terms in linguistics*. London: Routledge.

Tsvetkov, Yulia, Dyer, Chris, Levin, Lori and Bhatia, Archna (2013). Generating English determiners in phrase-based translation with synthetic translation options. In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia (Bulgaria), 271–280.

Ture, Ferhan, Oard, Douglas W. and Resnik, Philip (2012). Encouraging consistent translation choices. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal (Canada), 417–426.

Uryupina, Olga (2006). Coreference resolution with and without linguistic knowledge. In: *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC-2006)*, Genoa (Italy), 893–898.

Versley, Yannick, Ponzetto, Simone Paolo, Poesio, Massimo, Eidelman, Vladimir, Jern, Alan, Smith, Jason, Yang, Xiaofeng and Moschitti, Alessandro (2008). BART: A modular toolkit for coreference resolution. In: *Proceedings of the ACL-08: HLT Demo Session*, Columbus (Ohio, USA), 9–12.

Veselovská, Kateřina, Ngụy Giang Linh and Novák, Michal (2012). Using Czech-English parallel corpora in automatic identification of *it*. In: *Proceedings of the 5th Workshop on Building and Using Comparable Corpora*, Istanbul (Turkey), 112–120.

Voigt, Rob and Jurafsky, Dan (2012). Towards a literary machine translation: The role of referential cohesion. In: *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, Montréal (Canada), 18–25.

Wandmacher, Tonio and Antoine, Jean-Yves (2007). Methods to integrate a language model with semantic information for a word prediction component. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague (Czech Republic), 506–513.

Wäschle, Katharina and Riezler, Stefan (2012). Structural and topical dimensions in multi-task patent translation. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon (France), 818–828.

Witten, Ian H. and Bell, Timothy C. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37 (4):1085–1094.

Wong, Billy T. M. and Kit, Chunyu (2012). Extending machine translation evaluation metrics with lexical cohesion to document level. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island (Korea), 1060–1068.

Wong, Billy T. M., Pun, Cecilia F. K., Kit, Chunyu and Webster, Jonathan J. (2011). Lexical cohesion for evaluation of machine translation at document level. In: *Proceedings of the 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, Tokushima (Japan), 238–242.

Xiao, Tong, Zhu, Jingbo, Yao, Shujie and Zhang, Hao (2011). Document-level consistency verification in machine translation. In: *MT Summit XIII: the Thirteenth Machine Translation Summit*, Xiamen (China), 131–138.

Xiong, Deyi, Ding, Yang, Zhang, Min and Tan, Chew Lim (2013a). Lexical chain based cohesion models for document-level statistical machine translation. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle (Washington, USA), 1563–1573.

Xiong, Deyi, Guosheng, Ben, Zhang, Min, Lü, Yajuan and Liu, Qun (2013b). Modeling lexical cohesion for document-level machine translation. In: *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, Beijing (China), 2183–2189.

Xiong, Deyi and Zhang, Min (2013). A topic-based coherence model for statistical machine translation. In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, Bellevue (Washington, USA), 977–983.

Žabokrtský, Zdeněk, Ptáček, Jan and Pajas, Petr (2008). TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In: *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus (Ohio, USA), 167–170.

Zeman, Daniel (2010). Hierarchical phrase-based MT at the Charles University for the WMT 2010 shared task. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala (Sweden), 212–215.

Zhao, Bing and Xing, Eric P. (2006). BiTAM: Bilingual topic admixture models for word alignment. In: *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, Sydney (Australia), 969–976.

Zhao, Bing and Xing, Eric P. (2008). HM-BiTAM: bilingual topic exploration, word alignment, and translation. In: J. C. Platt, D. Koller, Y. Singer and S. Roweis (eds.), *Advances in Neural Information Processing Systems 20*, Cambridge (Mass.): MIT Press, 1689–1696.

ACTA UNIVERSITATIS UPSALIENSIS

Studia Linguistica Upsaliensia

Editors: Joakim Nivre and Åke Viberg

1. *Jörg Tiedemann*, Recycling translations. Extraction of lexical data from parallel corpora and their application in natural language processing. 2003.
2. *Agnes Edling*, Abstraction and authority in textbooks. The textual paths towards specialized language. 2006.
3. *Åsa af Geijerstam*, Att skriva i naturorienterande ämnen i skolan. 2006.
4. *Gustav Öquist,* Evaluating Readability on Mobile Devices. 2006.
5. *Jenny Wiksten Folkeryd,* Writing with an Attitude. Appraisal and student texts in the school subject of Swedish. 2006.
6. *Ingrid Björk,* Relativizing linguistic relativity. Investigating underlying assumptions about language in the neo-Whorfian literature. 2008.
7. *Joakim Nivre, Mats Dahllöf and Beáta Megyesi,* Resourceful Language Technology. Festschrift in Honor of Anna Sågvall Hein. 2008.
8. *Anju Saxena & Åke Viberg*, Multilingualism. Proceedings of the 23rd Scandinavian Conference of Linguistics. 2009.
9. *Markus Saers*, Translation as Linear Transduction. Models and Algorithms for Efficient Learning in Statistical Machine Translation. 2011.
10. *Ulrika Serrander*, Bilingual lexical processing in single word production. Swedish learners of Spanish and the effects of L2 immersion. 2011.
11. *Mattias Nilsson*, Computational Models of Eye Movements in Reading : A Data-Driven Approach to the Eye-Mind Link. 2012.
12. *Luying Wang,* Second Language Acquisition of Mandarin Aspect Markers by Native Swedish Adults. 2012.
13. *Farideh Okati*, The Vowel Systems of Five Iranian Balochi Dialects. 2012.
14. *Oscar Täckström,* Predicting Linguistic Structure with Incomplete and Cross-Lingual Supervision. 2013.
15. *Christian Hardmeier,* Discourse in Statistical Machine Translation. 2014.