Evaluation of Translation Technology

Walter Daelemans
University of Antwerp

Véronique Hoste
University College Ghent/ University of Ghent

Lacking widely accepted and reliable evaluation measures, the evaluation of Machine Translation (MT) and translation tools remains an open issue. MT developers focus on automatic evaluation measures such as BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) which primarily count n-gram overlap with reference translations and which are only indirectly linked to translation usability and quality. Commercial translation tools such as translation memories and translation workbenches are widely used and their developers claim usefulness in terms of productivity, consistency or quality. However, these claims are rarely proven using objective comparative studies. This collection dissects the state of the art in translation technology and translation tool development and provides quantitative and qualitative answers to the question how useful translation technology is.

Evaluation of translation technology requires a multifaceted approach. It involves the evaluation of the textual output quality in terms of intelligibility, accuracy, fidelity to its source text, and appropriateness of style and register. But it also takes into account the usability of supportive tools for creating and updating dictionaries, for post-editing texts, for controlling the source language, for customization of documents, for extendibility to new languages and for domain adaptability, etc. Finally, evaluation involves contrasting the costs and benefits of translation technology with those of human translation performance.

This collection comprises 10 original contributions from researchers and developers in the field. The volume is divided into two parts. The first addresses evaluation of Machine Translation, the second evaluation of Translation Tools.

Part I opens with an invited position paper of Andy Way (*A critique of statistical machine translation*) in which he analyzes the divide between on the one hand the developers of Statistical Machine Translation (SMT) systems, and on the other hand translators. In spite of the technical success of SMT, with phrase-based SMT dominating research and development, translators largely ignore it. According to Andy Way, the reason for this is the fact that the approach is perceived as being extremely difficult to understand, as its proponents are not interested in addressing any community other than their own. After a fascinating account of the early history of

SMT, the author argues convincingly that SMT has much to learn from other paradigms, including more linguistically sophisticated ones. He also criticizes the danger of over-optimizing systems when using only automatic MT evaluation methods.

The topic of evaluation methodology is further taken up by Paula Estrella, Andrei Popescu-Belis, and Maghi King (*The FEMTI guidelines for contextual MT evaluation: principles and resources*) in their introduction to the Framework for the Evaluation of Machine Translation in ISLE (FEMTI). This methodology takes into account the context of the use of an MT system and is based on ISO/IEC standards and guidelines for software evaluation. The methodology provides support tools and helps users define contextual evaluation plans. Context in terms of tasks, users, and input characteristics indeed plays an all-important role in evaluation. The web-based FEMTI application allows evaluation experts to share and refine their knowledge about evaluation.

Despite the high correlations with human judgements (e.g. Zhang et al., 2004), automatic metrics such as BLEU and NIST do not necessarily result in an actual improvement in translation quality (Way, Callison-Burch et al., 2006). Furthermore, a limitation of current automatic scores developed within SMT is the fact that they give only a very general indication of translation quality. Both the article of Bogdan Babych and Anthony Hartley, and the contribution of Nora Aranberri-Monasterio and Sharon O'Brien focus on more fine-grained MT evaluation, aiming at a more thorough error analysis which can help MT developers to focus on problematic categories. Bogdan Babych and Anthony Hartley (*Automated error analysis for multiword expressions: using BLEU-type scores for automatic discovery of potential translation errors*) adapt the BLEU metric to allow for the detection of systematic mistranslations of multiword expressions (MWE), and also to create a priority list of problematic issues. Two aligned parallel corpora serve as the basis for their experiments and they experiment both with rule-based and statistical MT systems. They show that their approach allows for the discovery of poorly translated MWEs both on the source and target language side. Even more specific is the evaluation of output of rule-based MT systems when translating –ing forms by Nora Aranberri-Monasterio and Sharon O' Brien (*Evaluating RBMT output for –ing forms: a study of four target languages*). These forms have a reputation for being hard to translate into e.g. French, Spanish, German, and Japanese and are therefore frequently addressed in controlled language rules which seek to reduce the ambiguities in the source text in order to improve the machine translation output. For the evaluation of the translation quality of the -ing-form, the authors opted for a human evaluation and show that Systran, a rule-based MT system, obtains reasonable accuracy (over 70%) in translating this form. Due to the labour-intensive nature of human evaluation, they also assess the agreement between the human scores and automatic metrics such as NIST, GTM, etc. and show good correlations. The authors conclude on the basis of their experimental work that the problem of the -ing forms is

overstated and explore a few possibilities for further improving these results.

Part I closes with yet another perspective on the evaluation of Machine Translation: recipient evaluation. This study is another nice application of the context-based evaluation of MT. In order to determine the usefulness of MT as a cost-effective way of providing more material in the language of minorities, Lynne Bowker (*Can Machine Translation meet the needs of official language minority communities in Canada? A recipient evaluation.*) investigates the reception of MT in the Canadian context where bilingualism is officially legislated. The reception of MT output by the two studied Official Language Minority Communities (OLMCs) was investigated by presenting four translation versions, viz. human translations and raw, rapidly post-edited and maximally post-edited MT output to members of the two OLMCs. Bowker's study reveals that whereas (rapidly and maximally post-edited) MT output could be acceptable for information assimilation in cases where there is a lack of ability to understand the source text, only high-quality translations are acceptable for information dissemination where translation is seen as a means for preserving or promoting a culture. Another interesting finding was that the 'average' recipients are more open to MT output than language professionals.

Part II of this volume addresses the evaluation of computer-aided translation tools (see e.g. Bowker, 2002 for an introduction). These tools include Translation Memories (TM), (bilingual) terminology management software, monolingual authoring tools (spelling, grammar, style checking), workflow management tools etc. A first question to be answered is whether current state of the art tools are perceived as useful by translators, and how they can be improved. Iulia Mihalache (*Social and economic actors in the evaluation of translation technologies. Creating meaning and value when designing, developing and using translation technologies*) discusses the advantages for companies as well as for translators of encouraging public evaluation of tools in on-line communities, and develops evaluation criteria from the perspective of translators communities, making use of different technology adoption models. She also discusses the 'how' of evaluation: a more complete understanding of translation technologies evaluation criteria is obtained if translators' attitudes, perceptions and behaviours related to technologies are studied in a multidisciplinary way from sociological, economic, psychological, and cultural perspectives. Alberto Fernández Costales (*The role of computer assisted translation in the field of software localization*) analyzes the effectiveness of computer assisted translation tools in Localization, the adaptation of a product to a particular locale. By empirically studying the usability and reliability of a particular tool (Passolo) for localizing a program, insight is provided into how translation tools can alleviate some of the challenges of localization. Besides improving text consistency and terminological coherence (but see Miguel Jiménez-Crespo's paper for contradictory results), the main advantage is that these

tools can save time, and thereby improve the productivity of localization experts.

Possible improvements in current Translation Memory technology are studied in the article of Lieve Macken (*In search of recurrent units of translation*). Translation Memories are currently sentence-based. This means that new text to be translated can only be matched with sentence-like segments, leading to limited recall in many cases. However, the number of matches can be increased if input is allowed to match sub-sentential segments. In a series of experiments, the degree of repetitiveness of different text types is compared, and performance of a sentential Translation Memory system is compared with a sub-sentential one. The results show that whereas sub-sentential memory systems are certainly a move in the right direction, they also sometimes lead to distracting translation suggestions. For solving the latter problem, better word alignment algorithms are necessary.

TM tools have changed the nature of translation by imposing a number of technological constraints that can in principle lead to either positive results (increased consistency) or negative results (increased decontextualization). Miguel Jiménez-Crespo (*The effect of translation memory tools in translated web texts: evidence from a comparative product-based study*) provides an empirical study on the often-debated question whether TMs improve or degrade translation quality. In a corpus-based study of 40,000 original and localized Spanish websites, he shows that the localized texts (translated using TMs) show higher numbers of inconsistencies at the typographic, lexical, and syntactic levels than spontaneously produced, non-translated texts, and therefore lead to lower levels of quality. While this article does not provide the last word in this discussion, it paves the way to interesting follow-up studies controlling for different variables that may influence the difference observed.

Acknowledgements

The authors would like to take this opportunity to thank all the authors for their contributions. The final contributions have undergone a detailed review followed by a thorough revision step. Our sincere thanks also go to the reviewers who helped us to assure the highest level of quality for this publication: Joost Buysschaert, Gloria Corpas Pastor, Alian Desilets, Andreas Eisele, Frederico Gaspari, David Farwell, Eva Forsbom, Johann Haller, David Langlois, Lieve Macken, Karolina Owczarzak, Jörg Tiedemann, Harold Somers. We also thank Aline Remael for her advice throughout the publication process and for some of the final formal editing with Jeremy Schreiber.

Bibliography

Bowker, L. (2002). Computer Aided Translation Technology: A Practical Introduction, University of Ottawa Press, Ottawa, Canada.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the Role of Bleu in Machine Translation Research. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp.249-256).Association for Computational Linguistics. Trento, Italy.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. *Proceedings of the Second Human Language Technologies Conference (HLT)* (pp.138-145). Morgan Kaufmann. San Diego, USA.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.J. (2002). BLEU: a method for automatic evaluation of Machine Translation. *Proceedings of the 40th Annual Metting of the Association for Computational Linguistics (ACL)* (pp.311-318). Association for Computational Linguistics. Philadelphia, USA.

Zhang, Y., Vogel, S. and Waibel, A. (2004). Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.(pp.2051-2055). European Language Resources Association. Lisbon, Portugal.

I EVALUATION OF MACHINE TRANSLATION

A critique of Statistical Machine Translation

Andy Way
Dublin City University

Phrase-Based Statistical Machine Translation (PB-SMT) is clearly the leading paradigm in the field today. Nevertheless—and this may come as some surprise to the PB-SMT community—most translators and, somewhat more surprisingly perhaps, many experienced MT protagonists find the basic model extremely *difficult to understand. The main aim of this paper, therefore, is to discuss why this might be the case. Our basic thesis is that proponents of PB-SMT do not seek to address any community other than their own, for they do not feel any* need *to do so. We demonstrate that this was not always the case; on the contrary, when statistical models of translation were first presented, the language used to describe how such a model might work was very conciliatory, and inclusive. Over the next five years, things changed considerably; once SMT achieved dominance particularly over the rule-based paradigm, it had established a position where it did not need to bring along the rest of the MT community with it, and in our view, this has largely pertained to this day. Having discussed these issues, we discuss three additional issues: the role of automatic MT evaluation metrics when describing PB-SMT systems; the recent syntactic embellishments of PB-SMT, noting especially that most of these contributions have come from researchers who have prior experience in fields other than statistical models of translation; and the relationship between PB-SMT and other models of translation, suggesting that there are many gains to be had if the SMT community were to open up more to the other MT paradigms.*

1. Introduction

It is clear that Phrase-Based Statistical Machine Translation (PB-SMT) is by far the most dominant paradigm in our field today, at least with respect to the research community. Younger readers, newcomers to the field and 'dyed in the wool' practitioners of SMT may not realise it, but 'twas not ever thus. In the early 1980s, the struggle to become *the* way that MT was done focused on two rule-based (RBMT) strategies: transfer (e.g. Arnold & des Tombe, 1987; Bennett & Slocum, 1985), and interlingual MT (e.g. (Carbonell et al.,1992; Rosetta, 1994). Even before SMT was propounded as a possible competitor to RBMT, Example-Based MT (EBMT) had been proposed in (Nagao, 1984).[1]

At the time SMT was proposed in Brown et al. (1988a), therefore, RBMT was the prevalent paradigm, which according to some researchers

was under pressure from EBMT, which "set out to supplant ('traditional') rule-based MT" (Nirenburg et al., 1993).

Interestingly, when we asked recently whether at the time he had really thought that EBMT could take over from RBMT, Nagao (1984) noted:

> Yes. I thought that RBMT had a limitation because we cannot write a complete grammar of analysis, transfer and generation consistently and completely, and that improving an RBMT system is quite difficult because no one is confident about what grammar rules are to be changed in what way to handle a particular expression, etc. in order to improve a system.In contrast, EBMT has a kind of a learning function by adding new translation pairs to handle new expressions. It is a very simple process.

With RBMT being dominant, and EBMT having had a few years' headstart, SMT was, therefore, truly the new kid on the block back in 1988.

In Section 2, we address the issue of the nature of the language used by the IBM team in seeking to put across their views, starting with the 1988 paper, and moving via Brown et al. (1990) and Brown et al. (1992) to Brown et al. (1993), perhaps *the* most cited paper on (S)MT even today. As an aide-memoire, we have taken the liberty of asking some of the MT protagonists of the time for their recollections of the presentations which accompanied some of these papers, and as a result contrast the content of the papers with the more provocative language used in the accompanying conference presentations. It is apparent that the MT community at the time were less than welcoming to the newcomers, and that the language used to purvey their displeasure regarding the proposed techniques was itself somewhat rich!

Nonetheless, at this juncture, it suffices to say that we believe there was a real sea change in the language used between the earliest paper of Brown et al. (1988a), and the well-known Computational Linguistics article of Brown et al. (1993). It is by no means surprising that Brown et al. (1992) was presented at a conference subtitled "Empiricist vs. Rationalist Methods in MT". That is, the tide was already turning at this point, and by the time the 1993 paper was published the SMT developers had (largely) won the day. From this point on, SMT was mainstream, and no longer had to appeal to the remainder of the MT community to justify its acceptance; if you couldn't keep up, you were left behind. We contend that this pertains right up to the present day, where for many PB-SMT is completely impenetrable.

Of course, when you are a member of any dominant group, you don't need to appeal to outsiders; you may choose to, or instead you may look inwardly and preach to the converted using a language only they understand. With respect to PB-SMT, it is by no means clear that today's protagonists are even *aware* that a sizeable community exists for whom their research is unintelligible; nor is it clear that even if they *did* know this that

they would necessarily know how to communicate their ideas to non-SMT people.[2]

We aim to make three further contributions in this paper. Firstly, in Section 3 we outline the basic models of SMT, followed by a short discussion of MT evaluation. In particular, given the influence of automatic MT evaluation metrics nowadays, we address the question as to whether the tail is wagging the dog; instead of looking at the automatic evaluation scores *per se*, not enough emphasis is placed on whether translation quality is *actually* increasing. We conclude this section by examining what's good and not so good when it comes to SMT.

Secondly, in Section 4 we provide comment on the recent syntactic embellishments of PB-SMT, noting especially that many of these contributions have come from researchers who have prior experience in fields other than statistical models of translation.

Thirdly, in Section 5 we relate PB-SMT to other models of translation. We expect this to be of interest not only to the previously mentioned constituencies, but also to the PB-SMT community itself, many of whom do not seem to be aware that there are indeed other ways, and a vast untapped literature for them to avail of; things are not necessarily 'novel' just because they've been 'discovered' in an SMT framework.

Finally, in Section 6 we conclude with what are, in our opinion, the lessons to be learnt by all of us as a community from our observations.

2. (In)accessibility of Statistical Models of MT

It is clear from the previous section that we are critical of how SMT researchers present their work. In this section, we set out our argument about why we believe appropriate explanations of today's mainstream statistical models of translation are currently lacking for the constituencies mentioned at the outset of this paper.

2.1. Adopting a Conciliatory Tone

In the original exposition of SMT (Brown et al., 1988a),[3] the language used in places is noteworthy for its conciliatory tone (our italics):

> "We wrote this *somewhat speculative* paper hoping to stimulate interest in applications of statistics to translation and to *seek cooperation* in achieving this difficult task,"
>
> "*the proposal may seem radical*"
>
> "Very little will be said about employment of conventional grammars. This omission ... may only reflect ... our uncertainty about the

> degree of grammar sophistication required. *We are keeping an open mind!*"
>
> "Not to interrupt the flow of *intuitive ideas*, we omit the discussion of the corresponding algorithms".

That is, in the written records, at any rate, the intention seems to be one of appealing to the remainder of the (mostly rule-based) community, pointing out that this is new and that they might not like this competing approach (the authors are aware of the "*many weighty objections to our ideas*"), but also that they would not be able to achieve the goal of high-quality translation without the help of the (mostly) linguistics-based experts already operating in the field. Perhaps most apposite for today's practitioners is the final sentence, where they state that they feel their results are hopeful "for future statistical translation methods *incorporating the use of appropriate syntactic structure information*" (cf. Chiang, 2005).[4]

If we are permitted a quick aside, it is remarkable that the paper was accepted at all for Coling in 1988. Peter Brown, first author on the IBM papers, kindly forwarded to us the original review of the paper that appeared as Brown et al. (1988b), which, despite his working for the past 14 years in statistical finance, still takes pride of place on his office wall:

Original Review of SMT for Coling 1988

> The validity of statistical (information theoretic) approach to MT has indeed been recognized, as the authors mention, by Weaver as early as 1949. And was universally recognized as mistaken by 1950. (cf. Hutchins, MT: Past, Present, Future, Ellis Horwood, 1986, pp 30 ff. and references therein).
>
> The crude force of computers is not science. The paper is simply beyond the scope of COLING.

Given the content of this review,[5] the programme chair Eva Hajičová (as well as, interestingly, Makoto Nagao, who as one of Eva's "five advisors" presumably was responsible for reading the MT abstracts) must have been at least a little reticent in allowing the paper to proceed to publication; given the situation today, we must as a community compliment them retrospectively on their open-mindedness in accepting the paper and helping kick-start the new paradigm.

Brown et al. (1988a) was presented as part of a panel at TMI on "Paradigms for MT", with contributions from Jaime Carbonell, Peter Brown, Victor Raskin and Harold Somers. The recollection of those present is very interesting. Pierre Isabelle recalls:

> Peter Brown is pretty good at being provocative and at TMI-88 he was at his best. If I remember correctly, he went as far as saying that

> statistical approaches were just about to eradicate rule-based MT research (the bread and butter of everyone except him in the room) in the same way it had already eradicated rule-based speech research. Peter Brown definitely made that particular statement in public, but I am not 100% sure it was at TMI-88. In any event, his talk at TMI did indeed start and end with hugely provocative statements (for the time). As for the technical substance of the talk, few if any people in the room were then in a position to understand it in any depth.
>
> We were all flabbergasted. All throughout Peter's presentation, people were shaking their heads and spurting grunts of disbelief or even of hostility.

Pierre goes on to say that the usual question and answer session was "a big mess", because:

> 1) Nobody had understood Peter's talk well enough to come up with technical questions or objections; and 2) in the heat of the moment, nobody was able to articulate the general disbelief into anything like a reasonable response to Peter's incredible statements.

Harold Somers, sitting next to Peter Brown on the panel, notes:

> My recollection is that he knew very well that people would be shocked, and his presentation was more "you ain't gonna like this but…".
>
> The audience reaction was either incredulous, dismissive or hostile. Someone probably said "Where's the linguistic intuition?" to which the answer would have been "Yes, that's the point, there isn't any".

Walter Daelemans recalls that "the Leuven Eurotra people weren't very impressed by the talk and laughed it away as a rehash of 'direct' (word-by-word) translation, which was probably a fair comment at the time."

With respect to the above comments, Peter Brown unsurprisingly has a somewhat different recollection of these early presentations:

> While it is my style to be provocative, a statement such as "eradicating rule-based MT" would not be provocative but simply antagonistic and that is not my style. I was very much aware that what we were saying would be controversial, and that our goal was to show that mathematically what we were doing was correct. However, what I believed then and believe today is that while there is an enormous role for linguistics in translation, the actual translation itself should be done in a mathematically coherent framework. Our goal was to present that framework and to show how far you could get with only

> minimal linguistics and then to excite people into imagining how far we could get with more linguistics incorporated in a mathematically coherent system.
>
> As for starting and ending with "hugely provocative" statements, my goal was to provoke debate and discussion not to be antagonistic. Trying to antagonize others just isn't my style. I checked with with my colleagues who attended the TMI conference and they agreed that it is just not something I would have done.

In order to better understand the tensions between competing paradigms at the time, ten Hacken (2001) contains a few appropriate observations regarding the climate around this period:

> In the research programme predominant at Coling 1988 a number of signs of a crisis can be recognized. *In MT, one of the main problems was that despite large-scale investment in terms of time and money, projects considered as state-of-the-art failed to produce solutions which could be used in actual practice. As far as MT was available, the technology it used was outdated*. (p.11, our italics)

Of course, what ten Hacken says about the lack of useable systems (it's fairly obvious that he's speaking about Eurotra here, partly from his own experience on the project) is completely true. However, today's rule-based proponents would take issue with the latter point, presumably.

He also, somewhat more controversially still, observes that "most of the MT researchers at Coling 1988 belonged to the first group", namely "a group of scientists who refuse to consider the problem seriously". Ten Hacken (2001) notes further that

> By the mid 1990s the crisis had reached such proportions that we even find an explicit description of it in (Melby, 1995). The tone of this work is highly pessimistic in the sense that *MT as it had been attempted for a long time was a hopeless enterprise and should be given up*. (*ibid*., our italics)

Of course, in the intervening period, it largely *has* been abandoned.

With respect to Brown et al. (1988b), ten Hacken (2001) includes them in "a group of scientists who explore the borderlines of the research programme in order to find out whether non-mainstream versions might be better" (*ibid*.). Their approach was in direct contrast to the common view of the time that "the obstacles to translating by means of the computer are primarily linguistic" (Lehrberger & Bourbeau, 1988, p.1).[6] Ten Hacken (2001) observes that already by 1998, "the statistical approach to MT ha(d) gained prominent status at the cost of the previously dominant linguistic approach" (p.2), so the 'non-mainstream' had very much become the *de*

facto standard. We address some of the reasons why statistical models of translation became so dominant in section 3.2 below.

2.2. Describing SMT for Non-Specialists

Moving on, the first of the two Computational Linguistics articles (Brown et al., 1990) introduces most of the terminology we all use today in SMT: word alignment, language and translation models, parameter estimation, decoding, fertility, distortion, and perplexity, among others. The ideas in this paper are decidedly more fully worked out, and far from omitting "complete mathematical specification(s) to a future report", as in the 1988 papers, this 1990 paper provides the basic equations necessary for SMT to be carried out. Nonetheless, these are explained in language likely to be understood by newcomers to probability theory. And again, the last sentence offers the hand of friendship to the rule-based practitioners: "We hope to... construct grammars for both French and English and to base future translation models on the grammatical constructs thus defined" (p. 84).

In Brown et al. (1992), the IBM team provide a somewhat weak attempt at casting SMT as transfer. Notwithstanding our evaluation of this paper, it is obvious that even trying to couch SMT in terms of transfer may appeal to rule-based protagonists. Indeed, Peter Brown's recollection is that "by that time people were interested in the statistical approach and were listening to the talk without simply writing it off as something completely inane". This is confirmed by Somers (2003, p. 323), who observes that by now SMT "was seen (by some) as a serious challenge to the by now traditional rule-based approach, this challenge typified by the (partly engineered) confrontational atmosphere at TMI-92 in Montreal".

It was in such an atmosphere that around this time, Fred Jelinek, the head of the IBM team, uttered his (in)famous remark "Every time I fire a linguist, my system's performance improves". While many people believe this to have been uttered in the context of MT, it seems instead to have been quoted in the area of speech recognition.[7] Nonetheless, it was clear that linguistic proponents of MT could see that they were next in the firing line.

2.3. Mathematical Formulation of SMT

Clearly by this time the tide was turning in favour of corpus-based models of translation, including EBMT. Somers notes (*ibid.*) that EBMT was also seen (as we have noted above) as a significant challenger to RBMT. Interestingly, Way (2009) notes that similar trends in the language used can be seen in EBMT papers at the time:

> At the very same conference where SMT was first proposed in Brown et al. (1988a), Sumita and Tsutsumi (1988) note that for them, two items for future work involved using "deeper analysis" and focussing on "rule acquisition". However, within three years,

> one of the same authors felt able to write that the fact that "EBMT has no rules" was one of the main advantages over RBMT (Sumita & Iida, 1991).

Nonetheless, given that the nature of EBMT sub-sentential alignments are more linguistically motivated than those of SMT, EBMT has remained more approachable to those of a less statistical bent (cf. the last paragraph before section 3.1 for other reasons why this might be the case).

Returning to SMT, in Brown et al. (1993), any pretence at staying in touch with the nonstatistical disappears completely. While they note that

> Today, the fruitful application of statistical methods to the study of machine translation is within the computational grasp of anyone with a well-equipped workstation," (Brown et al., 1993)

this is soon followed by:

> We assume the reader to be comfortable with Lagrange multipliers, partial differentiation, and constrained optimization as they are presented in a typical college calculus text, and to have a nodding acquaintance with random variables.

Of course, we are taking these quotes somewhat out of context, and the title of the paper by Brown et al. (1993) is, after all, "The mathematics of statistical machine translation: Parameter estimation". To provide a more balanced view, therefore, Peter Brown noted in a recent email conversation:

> As for the language, our goal was to explain what we were doing as clearly as possible. None of us had any background in linguistics, just like we have no background in finance, so we just wrote it using the language and terminology of statistics with which we are familiar. I imagine that were we to write a paper on finance today, some of the finance guys might complain about our terminology also. For what it's worth I think it's very important to get the mathematics straight when doing linguistics but once it is straight then linguistic knowledge will be what matters. In other words, it's not math *or* linguistics, but math *and* linguistics. Our goal was to establish the mathematical framework for MT so that the linguistically-minded could proceed with the research. I gather from your note that that has not happened and it's unfortunately either math guys or linguistic guys working on MT but not both working together.

Given the title and topic of the paper, it would be churlish to heap all the blame on the pioneering IBM group; indeed, one of the reasons why this paper is so well-regarded nowadays is that it's particularly clearly written. As a successful paper, perhaps Brown et al. (1993) was seen as *the* way to

put across ideas from the SMT community, rather than being just one way in which this innovative research could be communicated.

Whether this was done intentionally or not, it's true that from 1993 onwards, attempts to engage the established MT community had indeed fallen by the wayside, and certainly by the new millennium SMT had become the dominant paradigm with no incentive to engage with researchers from older/other paradigms.[8]

Finally, while we can accept Peter's words at face value, it's clear that neither the SMT community (at least not until recently, and only then when researchers from outside the mainstream SMT community started to demonstrate the effectiveness of syntax) nor the more linguistically-oriented researchers - who along with the linguists, have to take their fair share of the blame for allowing SMT to become so dominant despite the contents of these early SMT papers - took from the IBM research the fact that once the mathematics had been properly sorted out, "then linguistic knowledge will be what matters"; if they had, we'd probably have had ten years earlier the syntax-based systems that are coming onstream now.[9]

3. Phrase-Based Statistical Machine Translation

We have complained that papers on PB-SMT are somewhat less than perspicuous for the general MT audience. It's well outside the scope of this paper to try to explain the various components of such systems (corpus preparation, word alignment, phrase extraction, language and translation model induction, system tuning, decoding and post-processing) in a manner that is not overly loaded with terminology and formulae, and short on intuition. However, we will point the interested reader to two companion papers: firstly, in Hearne and Way (2009a), we *do* try to achieve exactly that, by providing an explanation of SMT for non-specialists; secondly, in Hearne and Way (2009b), we discuss the important role of translators and linguists in the SMT process, whose contribution is often overlooked by SMT developers, but nonetheless remains an absolute prerequisite for SMT as we know it today, as well as for any extensions going forward.

In a nutshell, the goal of PB-SMT is to find the most likely translation T of a source sentence S . We say "most likely," as many possible candidate target language translations may be proposed by the system. The most likely translation is the one with the highest probability (hence "argmax") according to $P(S \mid T).P(T)$, as in (1):

(1) $argmax_T \; P(S \mid T).P(T)$

where $P(S \mid T)$ is the *translation model*, which attempts to ensure that the meaning expressed in S is also captured in T , i.e. that T is an *adequate* translation of S ; and $P(T)$ is the *language model*, which tries to ensure that

the candidate translation *T* is actually a valid sentence in the target language, i.e. that *T* is *fluent*. This is the 'noisy channel' model of SMT (Brown et al., 1990; Brown et al., 1993), and the language and translation models are (usually) inferred from large monolingual and bilingual aligned corpora respectively.

It is commonplace today to use phrases rather than words as the basis of the statistical model (hence 'phrase-based'). A phrase is defined as a group of source words s□ that should be translated as a group of target words t□ The 'log-linear' model of PB-SMT (Och & Ney, 2002) (rather more flexible than the 'noisy channel' model) is that in (2):

(2) $argmax_T \sum_{m}^{M} {}_{=1} \lambda_m \cdot h_m (T, S)$

The uninitiated reader should note that the leftmost parts of the equations in (1) and (2) are identical, i.e. the task is the same; the only difference is how each candidate translation (out of the *T* possible translations) output by the SMT system is to be scored.

In (2), there are *M* feature functions, whose logarithms should be added together (hence the Σ in (2), as opposed to the multiplication in (1); the typical values for each feature are in practice so small that multiplying them becomes impractical, as the product of each of these probabilities approaches zero quite quickly) to give the overall score for each translation. Typical feature functions for a PB-SMT system include the phrase translation probabilities in both directions (i.e. source-to-target *P (t□ | s□)* and target-to-source *P(s□ | t□))* between the two languages in question, a target language model (exactly as in (1)), and some penalty scores to ensure that sentences of reasonable length vis-à-vis the input string are generated.[10] Note that if only the translation model and language model features were used, then the log-linear model in (2) would be identical to the noisy channel model in (1). Typically the λ_m weights for each feature h_m in (2) are optimized with respect to some particular scoring function (usually a specific evaluation metric, cf. section 3.1 for further discussion of this topic) on a development (or 'tuning') set using a technique called Mimimum Error Rate Training (MERT) (Och, 2003) to try to ensure optimal performance on a specific test set of sentences, which is hopefully as 'similar' as possible to the development set. We again refer interested readers to Hearne & Way (2009a) for more detailed description of the components of these models in language we hope is more intuitive to them than is usually seen in SMT papers.

In Hearne & Way (2009b), we make the following (hopefully useful) observation:

> RMBT and EBMT dwell on the process via which a translation is to be produced for each source sentence, whereas SMT dwells on how to tell which is the better of two or more proposed translations for a source sentence. Thus, RBMT and EBMT focus on the best way to

> generate a translation for each input string, whereas SMT focuses on generating many thousands of hypothetical translations for the input string and working out which one is most likely. In seeking to understand SMT in particular, this is a key distinction: while the means by which RBMT and EBMT generate translations usually look somewhat plausible to us humans, the methods of translation generation in SMT are not intuitively plausible. In fact, the methods used are not intended to be either linguistically or cognitively plausible (just probabilistically plausible) and holding onto the notion that they somehow are or should be simply hinders understanding of SMT.

Not everyone would agree with us regarding this latter point, and we return to this in section 3.3.

3.1. Evaluation

While we've excluded discussion of the pre-processing and runtime stages in PBSMT, one stage that warrants a few words here is evaluation.
Ten Hacken (2001) makes the following observation:

> Whereas the architecture of the system and the choice of a linguistic theory as a source of knowledge to be applied are the subject of controversial discussion, the assumptions on the nature of translation and the proper evaluation of the MT system are not questioned in the late 1980s (p.13).

We argue below that while the introduction of automatic evaluation metrics in MT - where MT system output is compared against one or more reference translations produced by humans - has largely been beneficial, they have to a large extent taken on too much importance, especially since real translation quality is what we should be concerned with.

In our view, today's automatic MT evaluation metrics are basically useful for three tasks:

1) for system developers to check that different incarnations of the *same* system are improving over time;
2) to compare *different* systems when trained and tested on the same data sets, as in today's large-scale MT evaluation campaigns such as NIST,[11] WMT[12] (Callison-Burch et al., 2007; Callison-Burch et al., 2008; Callison-Burch et al., 2009), IWSLT[13] (Paul, 2006; Fordyce, 2007; Paul, 2008) etc.;
3) for MERT (Och, 2003), i.e. customising (tuning) one's system to perform as well as it can on the current data using *one* particular MT evaluation metric (e.g. BLEU (Papineni et al., 2002), NIST (Doddington, 2002), WER (Levenshtein, 1966), Meteor (Banerjee & Lavie, 2005), F-Score (Turian et al., 2003), etc.).

In the original exposition of BLEU, the main use envisaged by such automatic evaluation metrics concerned the first task above, namely incremental testing of one particular system on a defined test set of example sentences. There is no doubt, especially on small-scale evaluation tasks (such as IWSLT, where only 20,000-40,000 training examples of parallel text are available), that these evaluation metrics are especially useful, as changes to the code base can be evaluated very quickly, and quite often.

What they are *not* so useful for is telling potential users which system is best for their purposes, i.e. if someone were considering purchasing an MT system, and wanted to know how to discern the performance of one system against the other, we would not necessarily advise their doing so on the basis of the systems' comparative BLEU scores. While that's exactly what's done in the second task above, users should realise that those scores represent the systems' scores trained on *one* data set for *one* language pair in *one* language direction and tested on *one* (small) set of sentences, all of which may or may not bear any relation to the *actual* scenario that the user has in mind in which the system is to be deployed. *Caveat emptor*!

With respect to the third scenario outlined above, there are any number of automatic evaluation metrics, from string-based (e.g. BLEU, NIST, F-Score, Meteor) to dependency-based (Liu & Gildea, 2005; Owczarzak et al., 2008). MERT is concerned with the optimisation of one's system performance to *one* such particular metric on a development set, and hoping that this carries forward to the test set at hand. What is *not* (usually) performed in this developmental phase is any examination as to whether increases in scoring with the particular automatic MT evaluation metric *actually* improve the output translations as measured by real users in real applications.

While most developers of MT evaluation metrics cite some correlation with human judgements, many real improvements in translation quality do not result in improved BLEU (or any other) score. For instance, consider the example in (3) from Hassan et al. (2009).

(3) **Source:** وأكد الجانبان على دور منظمة التجارة العالمية

Reference: *The two sides highlighted the role of the World Trade Organization*,

Baseline: *The two sides on the role of the World Trade Organization (WTO)*,

CCG: *The two parties reaffirmed the role of the World Trade Organization*,

Omitting verbs turns out to be a problem for baseline PB-SMT systems. Given the amount of morphological variants possible, the language model only has a few occurrences of each possible verbal inflected form in order to try to decide which output string it most prefers. Accordingly, very often it prefers an *n*-gram which does not contain *any* verb, as opposed to including a verb that has been observed only rarely. That is, with respect to (3), the baseline system prefers the bigram *sides on* over any combination with *sides* plus some (relevant) verb (*highlighted*, here).

Hassan et al. (2009) note that this is particularly the case when translating the notorious verbless Arabic sentences, as in (3); while the reference translation contains the verb *highlighted*, as there is no verb in the Arabic sentence it is hardly surprising that the baseline system outputs a translation with no verb. However, the system of Hassan et al. (2009), which incorporates supertags[14] (Bangalore & Joshi, 1999) from Combinatory Categorial Grammar (CCG: Steedman, 2000), contains a more grammatically strict language model than a standard word-level Markov model, and so exhibits a preference for the insertion of a verb with a similar meaning to that contained in the reference sentence. Nonetheless, as *reaffirmed* is not contained in the reference translation, this clear improvement in translation quality does not carry over to an improvement according to any string-based evaluation metric. However, in the recent IWSLT-07 evaluation, the supertag-based Arabic-English system described in Hassan et al. (2007) was adjudged to be ranked first by some margin in the human evaluation, despite this clear advantage of more readable output not carrying over to the automatic evaluation scores.

In sum, increased automatic evaluation scores do not necessarily reflect any actual improvements in translation quality. Furthermore, He and Way (2009) note that there is no guarantee that parameters tuned on one metric (e.g. BLEU) will lead to optimal translation scores on the same metric; rather, the score can be improved significantly by tuning on an entirely different metric (e.g. METEOR), especially where just a single reference translation is available (in WMT evaluation tasks, for instance). He and Way (2009) also observe that tuning on combinations of metrics can lead to more robust performance. Of course, for general-purpose MT systems, op-

timising settings via MERT cannot be done at all, given the lack of a standalone test set; rather, the system must be robust in the face of *any* user input. Considering all these factors, we believe that tuning one's system to a particular evaluation metric is very much a case of the tail wagging the dog, rather than the other way round. Automatic evaluation metrics continue to have their place, but in our view they have taken on rather too much significance, to the possible detriment of *real* improvements in translation quality.

3.2. What's *Good* about PB-SMT

While much of this paper is critical of a number of issues related to statistical models of translation, it would be altogether remiss of us if we were to avoid any mention of some of the benefits that PB-SMT has brought to the wider MT community. These include resources such as:

- *Sentence-aligned corpora*, e.g. Europarl (Koehn, 2005);
- *Tools* such as:
  - word and phrase alignment software (principally Giza++,[15] (Och & Ney, 2003));
  - language modelling toolkits (e.g. SRILM,[16] (Stolcke, 2002));
  - decoders (freely available, such as Pharaoh (Koehn, 2004), but more recently open-source, such as Moses[17] (Koehn et al., 2007));
  - evaluation software (e.g. BLEU (Papineni et al., 2002), NIST (Doddington, 2002), GTM (Turian et al., 2003), Meteor (Banerjee & Lavie, 2005).

In addition, as SMT is rooted in decision theory, it's absolutely clear why the system outputs a translation as the most probable, namely because that output string maximizes the product of the translation model $P(S \mid T)$ and the language model $P(T)$ in the noisy channel model (cf. (1)), or the joint probability of the target and source sentences in the log-linear equation (Och & Ney, 2002) (cf. (2), and section 3 above for more discussion).

It is also very clear that the evaluation campaigns (such as NIST, IWSLT, WMT etc.) have enabled systems to be compared against one another, as standard training, development and test data are made available for each campaign. As in other areas of language processing, this competitive edge has caused groups to try to improve their systems, and such campaigns have doubtlessly resulted in advances in the state of the art. However, Callison-Burch et al. (2006) demonstrate that using string-based evaluation metrics is decidedly unsuitable for comparing systems of quite different types (SMT vs. RBMT, say), which is why the ultimate arbiter of system performance in the WMT tasks remains human evaluation, although a

host of automatic evaluation scores are provided for each competing system.

3.3. What's *Less* Good about PB-SMT

For all these reasons, newcomers to the field of MT can very quickly build a system which is competitive compared to those systems of much more experienced groups in the field. Given the enormous ramp up in terms of resources needed, these resources (especially now that Moses is open-source) have been a huge help to newcomers to MT, as well as to more established groups.

However, in our view it remains to be seen whether PB-SMT is the leading method because it's the *best* way of doing MT,or because the tools *exist* which facilitate the rapid prototyping of systems on new language pairs, and different data sets.

While the provision of parallel training corpora (not just of use in SMT, of course) and decoders is very much appreciated by the community, one wonders how much we are now reliant on Philipp Koehn[18] coming up with more data sources and (open-source) software in order for the field to make further advances. For instance, it's not clear that enough is being done (a) to fix things that need fixing; and (b) to make any fixes which *have* been made available to the wider community.

As an example, consider the case of alignment templates (Och & Ney, 2004), which is quite closely related to the use of generalized templates in EBMT. As many others have shown (e.g. Brown, 1999; Cicekli & Güvenir, 2003; Way & Gough, 2003), the use of generalized templates can improve the coverage and quality of EBMT systems. Furthermore, researchers such as Maruyama and Watanabe (1992) stated that "there is no essential difference between translation examples and translation rules - translation examples are special cases of translation rules" (cf. section 2.3 for an alternative view at the time).

Nonetheless, quite clearly the use of alignment templates has not caught on in PB-SMT anywhere near as much as templates/rules in EBMT and RBMT.[19] This is not because they are not useful; Och and Ney (2004) demonstrated their utility several years ago. Rather, in our view it is simply because the developers of PB-SMT decoders have not (yet) made provision for their use in the code-base.

This is just like the situation with the use of phrases (cf. section 5) and syntax (cf. section 4.1) in other paradigms. Years before phrases and syntax were shown to be of benefit in PB-SMT, practitioners in RBMT and EBMT had been incorporating them into their systems;[20] from its inception (Nagao, 1984), EBMT has sought to translate new texts by means of a range of sub-sentential data (both lexical *and* phrasal) stored in the system's memory. As regards syntax, EBMT systems have been built using dependency trees (e.g. Watanabe, 1992; Menezes & Richardson, 2003), annotated constituency tree pairs (e.g. Hearne, 2005;, Hearne & Way, 2006), and pairs

of attribute-value matrices (e.g. Way, 2003), among other methods. In much the same way, we contend that alignment templates will become incorporated into mainstream PB-SMT in the near future (cf. Zhao & Al-Onaizan, 2008) in hierarchical phrase-based MT (Chiang, 2005), at which point *everyone* will use them.

Finally, while it's clear that statistical models of translation are modelled on a well-defined decision problem, there is undoubtedly a lack of perspicuity in PB-SMT when it comes to explaining the data. Back in the bad old days of MT, ten Hacken (2001, p. 2) observed that most researchers "take linguistic phenomena as discussed in theoretical linguistics as a basis for the identification of topics in MT". While we agree that we never want to go back to that way of doing things, today's preoccupation with the size of one's BLEU score has gone too far in the opposite direction, so that most PB-SMT researchers would be unable to tell you whether their systems were able to cope with particular cases of 'hard' translational phenomena (e.g. headswitching, relation-changing, etc. See Hearne et al. (2007) for a recent example of what's possible using this tried and tested terminology, and even if they could, they would find it difficult to tell you how such constructions were handled.[21]

On a related point, Galley et al. (2006) state (our italics): "the broad statistical MT program is aimed at a wider goal than the conventional rule-based program - it seeks to *understand and explain human translation data*, and automatically learn from it."

This seems to us to be so far from the truth that it would not be recognised at all by people from outside SMT. For starters, there's an entire body of research dedicated to this - namely, corpus-based translation studies - which Galley et al.(2006) seem to have missed completely. As we stated in Hearne and Way (2009b) (cf. section 3 above), we believe there to be no linguistic or cognitive plausibility in the statistical model of translation. What's more, in our view a statistical approach is almost the *least* appropriate way to go about understanding and explaining human translation data.

4. Extending the Basic Model

Until very recently, it proved difficult to incorporate syntactic knowledge in order to obtain better quality translation output from PB-SMT systems on large benchmark test suites. Worse still, Koehn et al. (2003) demonstrated that adding syntactic constraints harmed the quality of their PB-SMT system.

4.1. Adding Syntax Helps PB-SMT

However, as we stated in the previous section, researchers have recently shown that the basic model of PB-SMT can be improved by the integration of syntax. The first paper to demonstrate this on a large benchmark translation task was (Chiang, 2005). However, his derived transduction grammar does not rely on any linguistic annotations or assumptions, so that the formal syntax induced is not linguistically motivated and does not necessarily capture grammatical preferences in the output target sentences.

More recently, Galley et al. (2006) and Marcu et al. (2006) present two similar extensions of PB-SMT systems with syntactic structure on the target language side. Both employ tree-to-string (so-called xRS) transducers, but their methods of acquiring the xRS rules and training them are different (cf. Hassan et al., 2009, for discussion of these differences).

In a different strand of work, other researchers have demonstrated that lexical syntax in the form of 'supertags' can be used to improve translation quality on a range of language pairs (Hassan et al., 2009) (cf. (3) and resultant discussion above).

4.2. Some Observations

It is evident that given the importance of statistical linguistic processing in NLP in general, many researchers have crossed over from statistical parsing to SMT, and these individuals have contributed enormously to syntactic models of SMT. This is a good thing, as until recently the parsing and MT communities have largely been distinct.

However, such researchers are themselves more likely to come from mathematical, statistical or computer science backgrounds, with much of the linguistics surfacing as annotated data. One could argue that they have been able to enter the field, and contribute to improvements in the area, because current SMT discourse is more accessible to them.

Nonetheless, the fact that syntax has been shown to be of use in PB-SMT is in stark contrast to prominent members of the community - albeit those with no linguistic background to speak of - stating in invited talks at recent large MT gatherings that integrating syntax would not be beneficial, and that linguists and translators had no role to play in the development of today's state-of-the-art MT systems. You don't have to think long to see how ironic this is, when SMT (and other corpus-based) systems are entirely dependent on parallel text generated by human translators (see Ozdowska et al., 2009, for investigation of the effect on translation quality of training SMT systems with such more or less appropriate sets of training data).

One might, therefore, hope that these statements have proven themselves to be ill-founded and have since been largely put to bed. However, more recently Zollmann et al. (2008) demonstrated on a range of Arabic-English tasks that the hierarchical model of Chiang (2005) and the syntax-augmented model of Zollmann and Venugopal (2006) do not show consis-

tent improvements over a baseline PB-SMT system which is allowed access to reorderings up to 12 words apart, so perhaps the debate will continue for a while yet.

5. PB-SMT and other Models of Translation

At the time of writing, statistical models of MT have been around for 20 years, but MT in general has been around for much longer.

In the previous sections, we noted that syntax had been integrated into models of RBMT and EBMT long before showing itself to be of use in PB-SMT, and even here, most of the breakthroughs have come about from those MT researchers with a broader NLP background.

Furthermore, we predicted that the use of templates/rules, long since useful in EBMT and RBMT, will, as Och and Ney (2004) demonstrated, but which has not led to widespread adoption in SMT so far, prove beneficial in phrase-based models of translation also (cf. Zhao & Al-Onaizan, 2008, for a first step in this direction for tree-based models).

Even here, though, if one consults the list of references in Och and Ney (2004), not *one* EBMT or RBMT citation is seen. Prior to Marcu and Wong (2002), the primary *modus operandi* in SMT was word-based (Brown et al., 1990; Brown et al., 1993). That graduating to phrase-based models led to improvements in quality is unsurprising given that from the very beginning (Nagao, 1984), EBMT has used both word and phrase alignments to translate new input strings. However, try to find *any* attributions in the SMT literature to EBMT and you'll (largely) be wasting your time.

The point is, of course, that the PB-SMT community is remarkably inward-looking. Again, this is due to its dominance in the field of MT; not only is it the case that many SMT people do not see the need to provide access to their work to non-specialists because they do not think they have anything to contribute, but also SMT practitioners feel that there is little to be gained from accessing the wider MT literature. Those of us not operating solely in the mainstream are forced to consult the primary SMT literature, as it constitutes by far the bulk of what is published in our field today. Accordingly, most EBMT and RBMT papers contain references from SMT. Regarding the situation pertaining at ACL-COLING 1998, already eleven years ago ten Hacken (2001, p. 15) stated that "(some) researchers still clinging to the old values [...] have included at least a token reference to the new (statistical) values in order to increase their chances of being accepted". In our view, rather than an act of 'tokenism', in most cases non-SMT practitioners need to relate their work to the mainstream statistical models of translation in order to have a reasonable chance of getting their papers published, given (a) the relative lack of published research in other

areas, and (b) the preponderance of SMT-trained reviewers of conference and journal submissions.

There is much to be learned by the SMT community from the other paradigms. It should be noted that novelties are not so just because they've been 'discovered' in an SMT paradigm. One such example is Chang and Toutanova (2007), who discuss the difficulties associated with projecting dependency trees from source to target sentences, without mentioning in the text the term *transfer*, nor referring to any such works in the bibliography. More recently, Lopez (2008) finds that "Translation by Pattern Matching" avoids the problem of computing unfeasibly large statistical models in PB-SMT by extracting from the bilingual training corpus stored in memory only those source phrases and their aligned target equivalents suitable for translating the current input string. This is an *exact* description of pretty much any EBMT system. To be fair, Lopez (2008) does cite one EBMT paper, but the steps taken to avoid the term 'EBMT' are remarkable.
In Way (2009), we observed that

> There has undoubtedly been a colossal move away from RBMT to more statistical methods, but now the pendulum is swinging back (slowly) in the opposite direction . . . As a community we are moving up the "Vauquois Pyramid" (Vauquois, 1968) just like people were trying to do in the old rule-based times, but eventually, we will doubtless still need more than can be inferred from "just looking at annotated text pairs".

If this is true, then SMT practitioners will have to take these comments on board if they do not want to be left behind, in much the same way that the linguistic proponents of MT were left behind by the SMT movement.

6. Conclusion

In this paper, we have argued that today's predominant MT paradigm is largely incomprehensible to translators, and more surprisingly, to many experienced MT protagonists who are not statistically trained. This is largely an artefact, we claim, of how PB-SMT practitioners have chosen to present their work (cf. Hearne & Way, 2009a, for a somewhat more accessible description of SMT).

We showed that this was not always the case; when the original IBM research was presented, the language used was much more inclusive. However, as SMT became *the* principal way of doing MT, this conciliatory tone soon changed, to the point where today many people who *want* to understand have been left *so* far behind that they feel that it is impossible to ever catch up. We expressed the view that linguists and translators have to share the blame in allowing the field to move almost entirely in the statistical di-

rection, especially when the seminal IBM papers very much left the door open for collaboration with the linguistic community.

However, in our view SMT researchers will soon have to alter their position, if the use of syntax (and later, once a further ceiling has been reached, semantics) *is* to become mainstream in today's models. These syntactic improvements have largely come about from those practitioners with a wider background than is the norm in SMT. Those without a linguistic background, then, appear to have two choices: (i) to attempt to include the linguists, so that they may be of help; or (ii) to continue to exclude linguists, while at the same time trying to make sense out of *their* writings.

We also discussed the overly important role nowadays of automatic evaluation metrics, to the exclusion of *actual* improvements in the translations output by our systems as measured by real users in real applications. The organisers of the WMT task, in particular, are to be applauded for maintaining human evaluation as the primary means by which translation quality is measured.

Finally, we have pointed out that there is much to be gained from consulting the research literature from the other MT paradigms. RBMT and EBMT practitioners have learnt much from SMT, and those communities will, we are certain, be very happy for SMT practitioners to learn from them also.

Acknowledgements

This work is partially funded by Science Foundation Ireland (http://www.sfi.ie) awards 05/IN/1732, 06/RF/CMS064 and 07/CE/I1142. Many thanks to Pierre Isabelle, Harold Somers and Walter Daelemans for providing their recollections regarding the early presentations of SMT, and to Makoto Nagao for his thoughts on the impact of EBMT on RBMT. We are especially grateful to Peter Brown for sharing with us the intentions of the IBM group when it came to clearly putting down their thoughts regarding the new paradigm, and for providing the first review of SMT for inclusion here. Finally, thanks to Mikel Forcada and Felipe Sánchez Martínez for comments on an earlier draft of this paper.

Bibliography

Arnold, D. & des Tombe, L. (1987). Basic theory and methodology in EUROTRA. In S. Nirenburg, (Ed.), *Machine translation: Theoretical and methodological issues* (pp. 114-135). Cambridge, UK: Cambridge University Press.

Banerjee, S. & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization (pp. 65-73); Ann Arbor, MI, USA..

Bangalore, S. & Joshi, A. (1999). Supertagging: An approach to almost parsing. *Computational Linguistics*, *25*(2), 237-265.

Bennett, W. & Slocum, J., (1985). The LRC machine translation system. *Computational Linguistics*, *11*, 111-121.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., & Roossin, P. (1988). A statistical approach to French/English translation. In Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (TMI 1988) (pages not numbered); Pittsburgh, PA, USA.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., & Roossin, P. (1988). A statistical approach to language translation. In Proceedings of the 12th International Conference on Computational Linguistics (Vol.1, pp. 71-76); Budapest, Hungary, August 22-27, 1988. Budapest: John von Neumann Society for Computing Sciences.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., & Roossin, P. (1990). A statistical approach to Machine Translation. *Computational Linguistics*, *16*(2), 79-85.

Brown, P., Della Pietra, S., Della Pietra, V., Lafferty, J., & Mercer, R. (1992). Analysis, statistical transfer, and synthesis in Machine Translation. In Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas (pp. 83-10); Montreal, QC, Canada.

Brown, P., Della Pietra, S., Della Pietra, V., & Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263-311.

Brown, R. (1999). Adding linguistic knowledge to a lexical example-based translation system. In Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (pp. 22-32); Chester, England.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J.. (2007). (Meta-)evaluation of machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation (pp. 136-158); Prague, Czech Republic.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2008). Further (meta-)evaluation of Machine Translation (pp. 70-106). In Proceedings of the Third Workshop on Statistical Machine Translation; Columbus, OH, USA.

Callison-Burch, C., Koehn, P., Monz, C., & Schroeder, J. (2009). Findings of the 2009 workshop on statistical Machine Translation. In Proceedings of the Fourth Workshop on Statistical Machine Translation (pp. 1-28); Athens, Greece.

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in Machine Translation research. In EACL-2006, 11th Conference of the European Association of Computational Linguistics, Proceedings of the Conference (pp. 249-256); Trento, Italy.

Carbonell, J., Mitamura, T., & Nyberg, E., 3rd (1992). The KANT perspective: A critique of pure transfer (and pure interlingua, pure statistics,...). In 4th International Conference on Theoretical and Methodological Issues in Machine Translation (pp. 225-235); Montreal, QC, Canada.

Chang, P-C. & Toutanova K,. (2007). A discriminative syntactic word order model for machine translation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (pp. 9-16); Prague, Czech Republic.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation (pp. 263-270). In 43rd Annual Meeting of the Association for Computational Linguistics; Ann Arbor, MI, USA.

Cicekli, I. & Güvenir, A.. (2003). Learning translation templates from bilingual translation examples. In M. Carl & A. Way (Eds.), *Recent advances in example-based machine translation* (pp. 255-286). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Clark, S. & Curran, J. (2004). The importance of supertagging for wide-coverage CCG parsing. In Proceedings of the 20th International Conference on Computational Linguistics (COLING'04) (pp. 282-288); Geneva, Switzerland.

Doddington, G. (2002). Automatic evaluation of MT auality using n-gram co-occurrence statistics. In Proceedings of Human Language Technology Conference 2002 (pp. 138-145); San Diego, CA, USA.

Dorr, B., Pearl, L., Hwa, R., & Habash, N. (2002). DUSTer: A method for unravelling cross-language divergences for statistical word-level alignment. In Proceedings of the 5th Conference of the Association for Machine Translation in the Americas (AMTA-02) (pp. 31-43); Berlin/Heidelberg: Springer-Verlag.

Fordyce, C. (2007). Overview of the IWSLT07 evaluation campaign. In Proceedings of the International Workshop on Spoken Language Translation (pp. 1-12); Trento, Italy.

Galley, M., Graehl, J., Knight, K., Marcu D., DeNeefe, S., Wang, W.,& Thayer, I. (2006). Scalable inference and training of context-rich syntactic models. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (pp. 961-968); Sydney, Australia.

Habash, N., B. Dorr & C. Monz. 2009. Symbolic-to-statistical hybridization: extending generation-heavy machine translation. Machine Translation 22(4) (in press).

Hassan, H., Ma, Y., & Way, A. (2007). MaTrEx: The DCU machine translation system for IWSLT 2007. In Proceedings of the International Workshop on Spoken Language Translation (pp. 69-75); Trento, Italy.

Hassan, H., Sima'an, K. & Way, A. (2009). Syntactically lexicalized phrase-based SMT. *IEEE Transactions on Audio, Speech and Language Processing, 16* (7), 1260-1273.

He, Y. & Way, A. (2009). Improving the objective function in minimum error rate training. In Proceedings of the Twelfth Machine Translation Summit (pp. 238-245); Ottawa, Canada.

Hearne, M. (2005). *Data-oriented models of parsing and translation*. Ph.D. thesis, Dublin City University, Dublin, Ireland.

Hearne, M., Tinsley, J., Zhechev, V., & Way. A. (2007). Capturing translational divergences with a statistical tree-to-tree aligner. In Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007) (pp. 85-94); Skövde, Sweden.

Hearne, M. & Way, A. (2006). Disambiguation strategies for data-oriented translation. In Proceedings of the 11th Annual Conference of the European Association for Machine Translation (pp. 59-68); Oslo, Norway.

Hearne, M. & Way, A. (2009). Statistical machine translation: A guide for linguists and translators. COMPASS (in press).

Hearne, M. & Way, A. (2009). On the role of translations in state-of-the-art statistical machine translation. COMPASS (in press).

Hutchins, W. (1986). Machine Translation: past, present, future. Chichester, UK: Ellis Horwood. http://www.hutchinsweb.me.uk/PPF-2.pdf

Koehn, P. (2004). Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Machine translation: From real users to research*. Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (pp. 115-124). AMTA 2004, LNAI 3265. Berlin/Heidelberg: Springer-Verlag.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In Proceedings of Machine Translation Summit X (pp. 79-86); Phuket, Thailand.

Koehn, P. et al. (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of the ACL 2007 Demo and Poster Session (pp. 177-180); Prague, Czech Republic.

Koehn, P., Och, F., & Marcu, D. (2003). Statistical Phrase-Based Translation. In Proceedings of the Joint Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL) (pp. 127-133); Edmonton, AB, Canada.

Lehrberger, J. & Bourbeau, L. (1988). Machine Translation: Linguistic characteristics of MT systems and general methodology of evaluation. Amsterdam: John Benjamins.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*, 707-710.

Liu, D. & Gildea, D. (2005). Syntactic features for evaluation of machine translation. In Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (pp. 25-32); Ann Arbor, MI, USA.

Lopez, A. (2008). Tera-scale translation models via pattern matching. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008) (pp. 505-512); Manchester, UK.

Marcu, D., Wang, W., Echihabi, A., & Knight, K. (2006). SPMT: Statistical machine translation with syntactified target language phrases. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006) (pp. 44-52); Sydney, Australia.

Marcu, D. & Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-02) (pp. 133-139); Philadelphia, PA, USA.

Maruyama, H. & Watanabe, H. (1992). Tree cover search algorithm for example-based translation. In Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in MT, TMI-92 (pp. 173-184); Montréal, QC, Canada.

Melby, A. (1995). The possibility of language: A discussion of the nature of language, with implications for human and machine translation. Amsterdam: John Benjamins.

Menezes, A. & Richardson, S. (2003). A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In M. Carl & A. Way (Eds.), *Recent advances in example-based machine translation* (pp. 421-442). Dordrecht: Kluwer Academic Publishers.

Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn & R. Banerji (Eds.), *Artificial and human intelligence* (pp. 173-180. Amsterdam: North-Holland.

Nirenburg, S., Domashnev, C. & Grannes, D. (1993). Two approaches to matching in example-based machine translation. In Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation TMI '93: MT in the Next Generation (pp. 47-57); Kyoto, Japan.

Och, F. (2003). Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (pp. 160-167); Sapporo, Japan.

Och, F. & Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 295-302); Philadelphia, PA, USA.

Och, F. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19-51.

Och, F. & Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, *30*(4), 417-449.

Owczarzak, K., van Genabith J., & Way, A. (2008). Evaluating machine translation with LFG dependencies. *Machine Translation*, *21*(2), 95-119.

Ozdowska, S. & Way, A. (2009). Optimal bilingual data for French-English PB-SMT. In Proceedings of EAMT-09, the 13th Annual Meeting of the European Association for Machine Translation (pp. 96-103); Barcelona, Spain.

Papineni, K., Roukos, S., Ward, T., & Zhu, W-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02) (pp. 311-318); Philadelphia, PA, USA.

Paul, M. (2006). Overview of the IWSLT 2006 evaluation campaign. In Proceedings of the International Workshop on Spoken Language Translation (pp. 1-15); Kyoto, Japan.

Paul, M. (2008). Overview of the IWSLT 2008 evaluation campaign. In Proceedings of the International Workshop on Spoken Language Translation (pp. 1-17); Honolulu, HI, USA.

Rosetta, M. (1994). Compositional translation. Dordrecht: Kluwer Academic Publishers.

Sánchez Martínez, F. (2008). *Using unsupervised corpus-based methods to build rule-based machine translation systems*. Ph.D thesis, Universitat d'Alacant, Alacant, Spain.

Shannon, C., & Weaver, W. (1949). The mathematical theory of communication. Urbana, IL: University of Illinois Press.

Somers, H. (2003). Introduction to Part III: System Design. In S. Nirenburg, H. Somers & Y. Wilks (Eds.), *Readings in machine translation* (pp. 321-324). Cambridge, MA: The MIT Press.

Steedman, M. (2000). The syntactic process. Cambridge, MA: The MIT Press.

Stolcke, A. (2002). SRILM - An extensible language modeling toolkit. In Proceedings of the 7th International Conference on Spoken Language Processing (pp. 901-904); Denver, CO.

Sumita, E., & Iida, H. (1991). Experiments and prospects of example-based machine translation. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91), (pp. 185-192); Berkeley, CA, USA.

Sumita, E., & Tsutsumi, Y. (1988). A translation aid system using flexible text retrieval based on syntax-matching. In Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (TMI 1988), Proceedings Supplement, (pages not numbered); Pittsburgh, PA, USA.

ten Hacken, P. (2001). Has there been a revolution in machine translation? *Machine Translation 16*(1), 1-19.

Turian, J., Shen, L., & Melamed, D. (2003). Evaluation of machine translation and its evaluation. In Proceedings of Machine Translation Summit IX (pp. 386-393); New Orleans, LA, USA.

Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in machine translation. In IFIP Congress-68 (pp. 254-260); Edinburgh. Reprinted in C. Boitet (Ed.), *Bernard Vauquois et la TAO: Vingtcinq ans de traduction automatique – analectes* (pp. 201-213) (1988). Grenoble: Association Champollion.

Wahlster, W. (Ed.). (2000). Verbmobil: Foundations of speech-to-speech translation. Berlin: Springer-Verlag.

Watanabe, H. (1992). A similarity-driven transfer System. In Proceedings of the fifteenth (sic) International Conference on Computational Linguistics, COLING-92 (pp. 770-776); Nantes, France.

Way, A. (2003). Machine translation using LFG-DOP. In R. Bod, R. Scha & K. Sima'an (Eds.), *Data-Oriented Parsing* (pp. 359-384). Stanford, CA: Center for the Study of Language and Information.

Way, A. (2009.) Panning for EBMT gold, or "Remembering not to forget": The DCU Experience. *Machine Translation* (in press).

Way, A., & Gough, N. (2003). wEBMT: Developing and validating an EBMT system using the World Wide Web. *Computational Linguistics 29*(3), 421-457.

Zhao, B., & Al-Onaizan, Y. (2008). Generalizing local and non-local word-reordering patterns for syntax-based machine translation. In Proceedings of EMNLP 2008, Conference on Empirical Methods in Natural Language Processing, (pp. 572-581); Waikiki, HI, USA.

Zollmann, A., & Venugopal, A.. (2006). Syntax-augmented machine translation via chart parsing. In Proceedings of the Workshop on Statistical Machine Translation, HLT-NAACL (pp. 138-141); New York, NY, USA.

Zollmann, A., Venugopal, A., Och, F., & Ponte, J. (2008). A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008) (pp. 1145-1152); Manchester, UK.

---

[1] Although Nagao's paper dates from 1984, its contents were delivered in a presentation in 1981.

[2] Notable exceptions are Philipp Koehn and Kevin Knight, who have given many lucid tutorials on SMT at various conferences.

[3] This paper was followed ten weeks later by (Brown et al., 1988b). Note that (ten Hacken, 2001) incorrectly observes that (Brown et al., 1988b) was "probably the first presentation of the groundbreaking IBM project". Apart from the slightly different titles (cf. also the similarity of the title in (Brown et al., 1990)), the content of the papers barely differs. One (probably!) wouldn't get away with this nowadays.

[4] However, a portent of what was to come is the observation that "preliminary experiments... indicate that only a very crude grammar may be needed". See section 2.3 for more on this topic.

[5] If one consults (Hutchins, 1986), as the reviewer invites us to do, one notes, for example, that "Weaver's own favoured approach, the application of cryptanalytic techniques, was immediately recognised as mistaken" (section 2.4.1). However, Weaver also expounded the virtues of "the probabilistic foundations of communication theory" (as (Hutchins, 1986) puts it), so while it was right to say that the cryptanalytic approach was mistaken, it was far from correct to say that the ideas of (Shannon & Weaver, 1949) had no potential for application in MT.

[6] Interestingly, this is perhaps more true nowadays than it was 20 years ago! See section 4 for more discussion. Note that as has been made plain here, the dichotomy used by (ten Hacken, 2001) to explain the various approaches was not one shared by the IBM team. Rather, in their view, linguistic insight would be necessary once the model had been given an adequate mathematical description.

[7] This is confirmed by Peter Brown, who informed us that "Jelinek is famous for that statement and made it many times but with regard to speech recognition not translation." See http://en.wikiquote.org/wiki/Fred\Jelinek, where one particular source is given as a Workshop on Evaluation of NLP Systems, Wayne, PA, USA, December, 1988. Note that (ten Hacken, 2001) erroneously attributes this quote to Peter Brown (p.10).

[8] As a brief aside, around 1996 the IBM SMT team broke up, and went to work for Renaissance Technologies applying their statistical models to predict stock market fluctuations. Fortunately, around the same time, Hermann Ney took on four PhD students in Aachen—Franz Och, Stephan Vogel, Cristoph Tillmann, and Sonja Nießen—and Alex Waibel also took on YeYi Wang as an SMT student in Karlsruhe/CMU, both as a result of their participation in the Verbmobil project (Wahlster, 2000). It is interesting to speculate about what would have happened to SMT if this fresh (and clearly significant) input had not come onstream at that time; it is possible that SMT would have disappeared from view, for a while at least.

[9] Furthermore, while the latter point regarding the void between the statistical and linguistic camps is largely true even today, we address it in more detail in section 4.

[10] As stated, most developers of PB-SMT systems, including this author, refer to the model in equation (2) somewhat loosely as the 'log-linear' model. This is, of course, not entirely accurate; rather, it is a method whereby linear combinations of logarithms of probabilities may be combined. Of course, when things like word and phrase penalties are used as feature functions, one can quickly see that not even this is strictly true.

[11] National Institute of Standards and Technology: http://www.nist.gov/speech/tests/mt/

[12] Workshop on Statistical Machine Translation. For the 2009 edition see http://www.statmt.org/wmt09/.

[13] International Workshop on Spoken Language Translation. For the 2008 edition see http://www.slc.atr.jp/IWSLT2008/.

[14] Without going into unnecessary detail, a supertag essentially describes lexical information such as the Part-of-Speech tag and subcategorisation information of a word.

[15] http://www.fjoch.com/GIZA++.html

[16] http://www.speech.sri.com/projects/srilm/

[17] http://www.statmt.org/moses/

[18] Philipp maintains a rich source of information on SMT at http://www.statmt.org.

[19] For a novel application, see (Sánchez Martínez, 2008) who uses PB-SMT alignment templates to bootstrap the acquisition of transfer rules in the open-source Apertium RBMT platform (http://www.apertium. org). If our comments in section 5 are accurate, given the title of this work, these interesting findings will remain largely undiscovered by the SMT community.

[20] For the uninitiated, many people have criticised the use of the term "phrase" to describe the basic units of translation in PB-SMT. We will not add to this here, but will merely note that the term as used in PB-SMT has a quite different meaning to that used in traditional linguistics.

[21] Note that in one particular corpus, (Dorr et al., 2002) report that 10.5% of Spanish sentences and 12.4% of Arabic sentences have at least one such translation divergence, while in another, divergences relative to English occurred in around one third of Spanish sentences. (Habash et al., 2009) observe that "there is often overlap among the divergence types… with the categorial divergence occurring almost every time that there is any other type of divergence".