Phrase Linguistic Classification and Generalization for Improving Statistical Machine Translation

Adrià de Gispert
TALP Research Center
Universitat Politècnica de Catalunya (UPC)
Barcelona
agispert@gps.tsc.upc.es

Abstract

In this paper a method to incorporate linguistic information regarding single-word and compound verbs is proposed, as a first step towards an SMT model based on linguistically-classified phrases. By substituting these verb structures by the base form of the head verb, we achieve a better statistical word alignment performance, and are able to better estimate the translation model and generalize to unseen verb forms during translation. Preliminary experiments for the English - Spanish language pair are performed, and future research lines are detailed.

1 Introduction

Since its revival in the beginning of the 1990s, statistical machine translation (SMT) has shown promising results in several evaluation campaigns. From original word-based models, results were further improved by the appearance of phrase-based translation models.

However, many SMT systems still ignore any morphological analysis and work at the surface level of word forms. For highly-inflected languages, such as German or Spanish (or any language of the Romance family) this poses severe limitations both in training from parallel corpora, as well as in producing a correct translation of an input sentence.

This lack of linguistic knowledge in SMT forces the translation model to learn different translation probability distributions for all inflected forms of nouns, adjectives or verbs ('vengo', 'vienes', 'viene', etc.), and this suffers from usual data sparseness. Despite the recent efforts in the community to provide models with this kind of information (see Section 6 for details on related previous work), results are yet to be encouraging.

In this paper we address the incorporation of morphological and shallow syntactic information regarding verbs and compound verbs, as a first step towards an SMT model based on linguistically-classified phrases. With the use of POS-tags and lemmas, we detect verb structures (with or without personal pronoun, single-word or compound with auxiliaries) and substitute them by the base form[1] of the head verb. This leads to an improved statistical word alignment performance, and has the advantages of improving the translation model and generalizing to unseen verb forms, during translation. Experiments for the English - Spanish language pair are performed.

The organization of the paper is as follows. Section 2 describes the rationale of this classification strategy, discussing the advantages and difficulties of such an approach. Section 3 gives details of the implementation for verbs and compound verbs, whereas section 4 shows the experimental setting used to evaluate the quality of the alignments. Section 5 explains the current point of our research, as well as both our most-immediate to-do tasks and our medium and long-term experimentation lines. Finally, sections 6 and 7 discuss related works that can be found in literature and conclude, respectively.

[1] The terms 'base form' or 'lemma' will be used equivalently in this text.

Proceedings of the ACL Student Research Workshop, pages 67–72,
Ann Arbor, Michigan, June 2005.

2 Morphosyntactic classification of translation units

State-of-the-art SMT systems use a log-linear combination of models to decide the best-scoring target sentence given a source sentence. Among these models, the basic ones are a translation model $Pr(e|f)$ and a target language model $Pr(e)$, which can be complemented by reordering models (if the language pairs presents very long alignments in training), word penalty to avoid favoring short sentences, class-based target-language models, etc (Och and Ney, 2004).

The translation model is based on phrases; we have a table of the probabilities of translating a certain source phrase $\tilde{f}_j$ into a certain target phrase $\tilde{e}_k$. Several strategies to compute these probabilities have been proposed (Zens et al., 2004; Crego et al., 2004), but none of them takes into account the fact that, when it comes to translation, many different inflected forms of words share the same translation. Furthermore, they try to model the probability of translating certain phrases that contain just auxiliary words that are not directly relevant in translation, but play a secondary role. These words are a consequence of the syntax of each language, and should be dealt with accordingly.

For examples, consider the probability of translating 'in the' into a phrase in Spanish, which does not make much sense in isolation (without knowing the following meaning-bearing noun), or the modal verb 'will', when Spanish future verb forms are written without any auxiliary.

Given these two problems, we propose a classification scheme based on the base form of the phrase head, which is explained next.

2.1 Translation with classified phrases

Assuming we translate from f to e, and defining $\tilde{e}_i$, $\tilde{f}_j$ a certain source phrase and a target phrases (sequences of contiguous words), the phrase translation model $Pr(\tilde{e}_i|\tilde{f}_j)$ can be decomposed as:

$$\sum_{T} Pr(\tilde{e}_i|T, \tilde{f}_j) Pr(\tilde{E}_i|\tilde{F}_j, \tilde{f}_j) Pr(\tilde{F}_j, \tilde{f}_j) \qquad (1)$$

where $\tilde{E}_i$, $\tilde{F}_j$ are the generalized classes of the source and target phrases, respectively, and $T = (\tilde{E}_i, \tilde{F}_j)$ is the pair of source and target classes used, which we call Tuple. In our current implementation, we consider a classification of phrases that is:

- *Linguistic*, ie. based on linguistic knowledge
- *Unambiguous*, ie. given a source phrase there is only one class (if any)
- *Incomplete*, ie. not all phrases are classified, but only the ones we are interested in
- *Monolingual*, ie. it runs for every language independently

The second condition implies $Pr(\tilde{F}|\tilde{f}) = 1$, leading to the following expression:

$$Pr(\tilde{e}_i|\tilde{f}_j) = Pr(\tilde{E}_i|\tilde{F}_j) Pr(\tilde{e}_i|T, \tilde{f}_j) \qquad (2)$$

where we have just two terms, namely a standard phrase translation model based on the classified parallel data, and an instance model assigning a probability to each target instance given the source class and the source instance. The latter helps us choose among target words in combination with the language model.

2.2 Advantages

This strategy has three advantages:

Better alignment. By reducing the number of words to be considered during first word alignment (auxiliary words in the classes disappear and no inflected forms used), we lessen the data sparseness problem and can obtain a better word alignment. In a secondary step, one can learn word alignment relationships inside aligned classes by realigning them as a separate corpus, if that is desired.

Improvement of translation probabilities. By considering many different phrases as different instances of a single phrase class, we reduce the size of our phrase-based (now class-based) translation model and increase the number of occurrences of each unit, producing a model $Pr(\tilde{E}|\tilde{F})$ with less perplexity.

Generalizing power. Phrases not occurring in the training data can still be classified into a class, and therefore be assigned a probability in the translation model. The new difficulty that rises is how to produce the target phrase from the target class and the source phrase, if this was not seen in training.

2.3 Difficulties

Two main difficulties[2] are associated with this strategy, which will hopefully lead to improved translation performance if tackled conveniently.

Instance probability. On the one hand, when a phrase of the test sentence is classified to a class, and then translated, how do we produce the instance of the target class given the tuple T and the source instance? This problem is mathematically expressed by the need to model the term of the $Pr(\tilde{e}_i|T, \tilde{f}_j)$ in Equation 2.

At the moment, we learn this model from relative frequency across all tuples that share the same source phrase, dividing the times we see the pair $(\tilde{f}_j, \tilde{e}_i)$ in the training by the times we see $\tilde{f}_j$.

Unseen instances. To produce a target instance $\tilde{f}$ given the tuple T and an unseen $\tilde{e}$, our idea is to combine both the information of verb forms seen in training *and* off-the-shelf knowledge for generation. A translation memory can be built with all the seen pairs of instances with their inflectional affixes separated from base forms.

For example, suppose we translate from English to Spanish and see the tuple T=(V[go],V[ir]) in training, with the following instances:

I will go PRP(1S) will VB	iré VB 1S F
you will go PRP(2S) will VB	irás VB 2S F
you will go PRP(2S) will VB	vas VB 2S P

[2] A third difficulty is the classification task itself, but we take it for granted that this is performed by an independent system based on other knowledge sources, and therefore out of scope here.

where the second row is the analyzed form in terms of person (1S: 1st singular, 2S: 2nd singular and so on) and tense (VB: infinitive and P: present, F: future). From these we can build a generalized rule independent of the person ' PRP(X) will VB ' that would enable us to translate 'we will go' to two different alternatives (present and future form):

we will go	VB 1P F
we will go	VB 1P P

These alternatives can be weighted according to the times we have seen each case in training. An unambiguous form generator produces the forms 'iremos' and 'vamos' for the two Spanish translations.

3 Classifying Verb Forms

As mentioned above, our first and basic implementation deals with verbs, which are classified unambiguously before alignment in training and before translating a test.

3.1 Rules used

We perform a knowledge-based detection of verbs using deterministic automata that implement a few simple rules based on word forms, POS-tags and word lemmas, and map the resulting expression to the lemma of the head verb (see Figure 1 for some rules and examples of detected verbs). This is done both in the English and the Spanish side, and before word alignment.

Note that we detect verbs containing adverbs and negations (underlined in Figure 1), which are ordered before the verb to improve word alignment with Spanish, but once aligned they are reordered back to their original position *inside* the detected verb, representing the real instance of this verb.

4 Experiments

In this section we present experiments with the Spanish-English parallel corpus developed in the framework of the LC-STAR project. This corpus consists of transcriptions of spontaneously spoken dialogues in the tourist information, appointment scheduling and travel planning domain. Therefore, sentences often lack correct syntactic structure. Preprocessing includes:

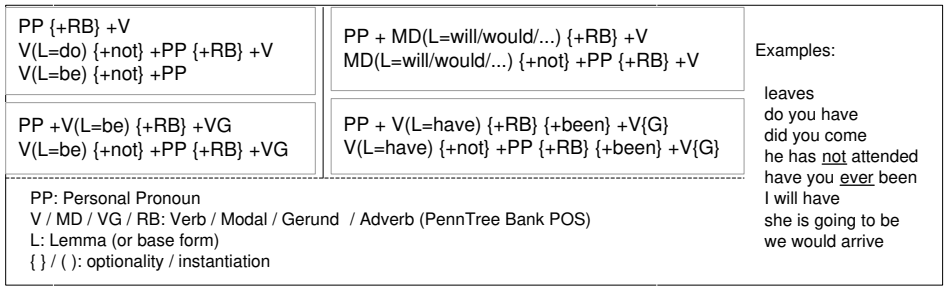

Figure 1: *Some verb phrase detection rules and detected forms in English.*

- Normalization of contracted forms for English (ie. wouldn't = would not, we've = we have)
- English POS-tagging using freely-available *TnT* tagger (Brants, 2000), and lemmatization using *wnmorph*, included in the WordNet package (Miller et al., 1991).
- Spanish POS-tagging using *FreeLing* analysis tool (Carreras et al., 2004). This software also generates a lemma or base form for each input word.

4.1 Parallel corpus statistics

Table 1 shows the statistics of the data used, where each column shows number of sentences, number of words, vocabulary, and mean length of a sentence, respectively.

	sent.	words	vocab.	Lmean
Train set				
English	29998	419113	5940	14.0
Spanish		388788	9791	13.0
Test set				
English	500	7412	963	14.8
Spanish		6899	1280	13.8

Table 1: *LC-Star English-Spanish Parallel corpus.*

There are 116 unseen words in the Spanish test set (1.7% of all words), and 48 unseen words in the English set (0.7% of all words), an expected big difference given the much more inflectional nature of the Spanish language.

4.2 Verb Phrase Detection/Classification

Table 2 shows the number of detected verbs using the detection rules presented in section 3.1, and the number of different lemmas they map to. For the test set, the percentage of unseen verb forms and lemmas are also shown.

	verbs	unseen	lemmas	unseen
Train set				
English	56419		768	
Spanish	54460		911	
Test set				
English	1076	5.2%	146	4.7%
Spanish	1061	5.6%	171	4.7%

Table 2: *Detected verb forms in corpus.*

In average, detected English verbs contain 1.81 words, whereas Spanish verbs contain 1.08 words. This is explained by the fact that we are including the personal pronouns in English and modals for future, conditionals and other verb tenses.

4.3 Word alignment results

In order to assess the quality of the word alignment, we randomly selected from the training corpus *350 sentences*, and a manual gold standard alignment has been done with the criterion of Sure and Possible links, in order to compute Alignment Error Rate (AER) as described in (Och and Ney, 2000) and widely used in literature, together with appropriately redefined Recall and Precision measures. Mathematically, they can be expressed thus:

$$recall = \frac{|A \cap S|}{|S|}, \quad precision = \frac{|A \cap P|}{|A|}$$

$$AER = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

where A is the hypothesis alignment and S is the set of Sure links in the gold standard reference, and P includes the set of Possible *and* Sure links in the gold standard reference.

We have aligned our data using GIZA++ (Och, 2003) from English to Spanish and vice versa (performing 5 iterations of model IBM1 and HMM, and 3 iterations of models IBM3 and IBM4), and have evaluated two symmetrization strategies, namely the union and the intersection, the union always rating the best. Table 3 compares the result when aligning words (current baseline), and when aligning classified verb phrases. In this case, after the alignment we substitute the class for the original verb form and each new word gets the same links the class had. Of course, adverbs and negations are kept apart from the verb and have separate links.

	Recall	Precision	AER
baseline	74.14	86.31	20.07
with class. verbs	76.45	89.06	**17.37**

Table 3: *Results in statistical alignment.*

Results show a significant improvement in AER, which proves that verbal inflected forms and auxiliaries do harm alignment performance in absence of the proposed classification.

4.4 Translation results

We have integrated our classification strategy in an SMT system which implements:

- $Pr(\tilde{e}_i|\tilde{f}_k)$ as a tuples language model (Ngram), as done in (Crego et al., 2004)
- $Pr(e)$ as a standard Ngram language model using SRILM toolkit (Stolcke, 2002)

Parameters have been optimised for BLEU score in a 350 sentences development set. Three references are available for both development and test sets. Table 4 presents a comparison of English to Spanish translation results of the baseline system and the configuration with classification (without dealing with unseen instances). Results are promising, as we achieve a significant mWER error reduction, while still leaving about 5.6 % of the verb forms in the test without translation. Therefore, we expect a further improvement with the treatment of unseen instances.

	mWER	BLEU
baseline	23.16	0.671
with class. verbs	22.22	0.686

Table 4: *Results in English to Spanish translation.*

5 Ongoing and future research

Ongoing research is mainly focused on developing an appropriate generalization technique for unseen instances and evaluating its impact in translation quality.

Later, we expect to run experiments with a much bigger parallel corpus such as the European Parliament corpus, in order to evaluate the improvement due to morphological information for different sizes of the training data. Advanced methods to compute $Pr(\tilde{e}_i|T, \tilde{f}_j)$ should also be tested (based on source and target contextual features).

The next step will be to extend the approach to other potential classes such as:

- Nouns and adjectives. A straightforward strategy would classify all nouns and adjectives to their base form, reducing sparseness.
- Simple Noun phrases. Noun phrases with or without article (determiner), and with or without preposition, could also be classified to the base form of the head noun, leading to a further reduction of the data sparseness, in a subsequent stage. In this case, expressions like `at night`, `the night`, `nights` or `during the night` would all be mapped to the class 'night'.
- Temporal and numeric expressions. As they are usually tackled in a preprocessing stage in current SMT systems, we did not deal with them here.

More on a long-term basis, ambiguous linguistic classification could also be allowed and included in the translation model. For this, incorporating statistical classification tools (chunkers, shallow parsers, phrase detectors, etc.) should be considered, and evaluated against the current implementation.

6 Related Work

The approach to deal with inflected forms presented in (Ueffing and Ney, 2003) is similar in that it also tackles verbs in an English – Spanish task. However, whereas the authors join personal pronouns and auxiliaries to form extended English units and do not transform the Spanish side, leading to an increased English vocabulary, our proposal aims at reducing both vocabularies by mapping all different verb forms to the base form of the head verb.

An improvement in translation using IBM model 1 in an Arabic – English task can be found in (Lee, 2004). From a processed Arabic text with all prefixes and suffixes separated, the author determines which of them should be linked back to the word and which should not. However, no mapping to base forms is performed, and plurals are still different words than singulars.

In (Nießen and Ney, 2004) hierarchical lexicon models including base form and POS information for translation from German into English are introduced, among other morphology-based data transformations. Finally, the same pair of languages is used in (Corston-Oliver and Gamon, 2004), where the inflectional normalization leads to improvements in the perplexity of IBM translation models and reduces alignment errors. However, compound verbs are not mentioned.

7 Conclusion

A proposal of linguistically classifying translation phrases to improve statistical machine translation performance has been presented. This classification allows for a better translation modeling and a generalization to unseen forms. A preliminary implementation detecting verbs in an English – Spanish task has been presented. Experiments show a significant improvement in word alignment, and in preliminary translation results. Ongoing and future research lines are discussed.

References

T. Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proc. of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.

X. Carreras, I. Chao, L. Padr´o, and M. Padr´o. 2004. Freeling: An open-source suite of language analyzers. *4th Int. Conf. on Language Resources and Evaluation, LREC'04*, May.

S. Corston-Oliver and M. Gamon. 2004. Normalizing german and english inflectional morphology to improve statistical word alignment. *Proc. of the 6th Conf. of the Association for Machine Translation in the Americas*, pages 48–57, October.

J.M. Crego, J. Mariño, and A. de Gispert. 2004. Finite-state-based and phrase-based statistical machine translation. *Proc. of the 8th Int. Conf. on Spoken Language Processing, ICSLP'04*, pages 37–40, October.

Y.S. Lee. 2004. Morphological analysis for statistical machine translation. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Short Papers*, pages 57–60, Boston, Massachusetts, USA, May. Association for Computational Linguistics.

G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, and R. Tengi. 1991. Five papers on wordnet. *Special Issue of International Journal of Lexicography*, 3(4):235–312.

S. Nießen and H. Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204, June.

F.J. Och and H. Ney. 2000. Improved statistical alignment models. *38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, October.

F.J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, December.

F.J. Och. 2003. Giza++ software. http://www-i6.informatik.rwth-aachen.de/˜och/ software/giza++.html.

A. Stolcke. 2002. Srilm - an extensible language modeling toolkit. *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, September.

N. Ueffi ng and H. Ney. 2003. Using pos information for smt into morphologically rich languages. *10th Conf. of the European Chapter of the Association for Computational Linguistics*, pages 347–354, April.

R. Zens, F.J. Och, and H. Ney. 2004. Improvements in phrase-based statistical machine translation. *Proc. of the Human Language Technology Conference, HLT-NAACL'2004*, pages 257–264, May.