Improving Statistical Machine Translation by Classifying and Generalizing Inflected Verb Forms

Adrià de Gispert, José B. Mariño and Josep M. Crego

TALP Research Center
Universitat Politècnica de Catalunya Barcelona
{agispert|canton|jmcrego}@gps.tsc.upc.es

Abstract

This paper introduces a rule-based classification of single-word and compound verbs into a statistical machine translation approach. By substituting verb forms by the lemma of their head verb, the data sparseness problem caused by highly-inflected languages can be successfully addressed. On the other hand, the information of seen verb forms can be used to generate new translations for unseen verb forms. Translation results for an English to Spanish task are reported, producing a significant performance improvement.

1. Introduction

Despite recent efforts to introduce linguistic information into Statistical Machine Translation (SMT) models [1], most of the current SMT systems are still ignoring morphologic analysis and work at the superficial level of word forms. For highly-inflected languages, such as Spanish (or any other language of the Romance family), this poses severe limitations both in training from parallel corpora, and in producing a correct translation of a test sentence. This is mainly due to the data sparseness caused by the translation model being forced to learn different probability distributions for all the inflected forms of verbs, nouns or adjectives.

Some previous research efforts to deal with this issue can be found in [2] and [3]. In the former work, the authors also tackle verbs in an English – Spanish task by joining personal pronouns and auxiliaries to form extended English units without transforming the Spanish side, which leads to an increased English vocabulary. On the other hand, in the latter the authors transform the text in the more-inflected language (Spanish) to separate base forms and suffixes for verb forms, improving performance when translating into English.

In this paper we address the incorporation of morphological and shallow-syntax information regarding verbs and compound verbs into an SMT system, as a first step towards a model based on linguistically-classified phrases. With the use of rules that incorporate POS-tags and lemmas, verb structures (with or without personal pronoun, single-word or compound with auxiliaries) are detected and substituted by the lemma of the head verb. This way, each inflected or compound form of a verb shares the same probability distribution, and a new instance model is proposed to help the decoder choose the adequate verb form given the source verb form. A novel generalization strategy for unseen verb forms is also introduced. Experiments for the English to Spanish translation direction (from a less inflected to a more inflected language) are reported, showing very promising results.

The paper is organized as follows. Section 2 discusses the motivation of this classification strategy and gives details of its implementation in a real SMT system. Special attention is given to the proposed instance model and generalization technique to deal with unseen verb forms. Section 3 introduces the parallel corpus used and reports translation results. Finally, Sections 4 discusses the approach and outlines future research lines to be explored.

2. Morpho-syntactic classification of translation units

State-of-the-art SMT systems use a log-linear combination of models to decide the best-scoring target sentence e given a source sentence f.

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\} \qquad (1)$$

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, lexical probability models, class-based target-language models, etc [4].

The translation model is usually based on phrases, that is, 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 [5, 6], 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 of verb forms based on the lemma of the head verb, which is explained next.

This work has been partially supported by the Spanish Government under grant TIC2002-04447-C02 (ALIADO project), the European Union under grant FP6-506738 (TC-STAR project) and the Dep. of Universities, Research and Information Society (Generalitat de Catalunya).

2.1. Translation with classified phrases

Suppose we want to translate a source sentence f to target sentence e. By defining $\tilde{e}_i$ as a certain source phrase and $\tilde{f}_j$ as a target phrase (where phrases are just sequences of contiguous words), the phrase translation model $Pr(\tilde{e}_i|\tilde{f}_j)$ can be decomposed as:

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

where $T = (\tilde{E}_i, \tilde{F}_j)$ is the pair of source and target classes used (called Tuple), and $\tilde{E}_i$, $\tilde{F}_j$ are the generalized classes of the source and target phrases, respectively. 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 $Pr(\tilde{F}_j|\tilde{f}_j) = 1$, leading to the following expression:

$$Pr(\tilde{e}_i|\tilde{f}_j) \approx \max_T Pr(\tilde{E}_i|\tilde{F}_j) Pr(\tilde{e}_i|T, \tilde{f}_j) \quad (3)$$

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. Instance model

In order to estimate this instance model $Pr(\tilde{e}_i|T, \tilde{f}_j)$, we propose a simple approach based on the relative frequency of each instance across all tuples that share the same source phrase, as expressed by equation 4.

$$Pr(\tilde{e}_i|T, \tilde{f}_j) = \frac{N(T, \tilde{e}_i, \tilde{f}_j)}{N(T, \tilde{f}_j)} \quad (4)$$

thus weighing each target verb form given the source form, and the translation tuple or phrase containing the source and target classes.

2.3. Generalization of unseen verb forms

Usually, a number of verb forms appearing in the test set will be unseen in the training data. In these cases, they will be classified to the lemma of their head verb and, if this has been seen, will be translated into a target phrase. However, the instance model probability given this source verb form is not defined, and a generalization strategy must be followed.

To produce a target instance $\tilde{e}_i$ given the tuple T and an unseen source instance $\tilde{f}_j$, the approach followed has been to make use of the information of verb forms that are seen in the training, seeking among seen instances those that are identical except on the personal pronoun (or verb suffix).

For example, suppose we want to translate the sentence '*we would have payed* it' from English to Spanish and we see tuples T_1=(V[pay],V[pagar]), T_2=T(V[pay],V[hacer] el pago) and T_3=T(V[pay] it, lo V[pagar]) translating the class V[pay] in the training data. However, among all seen instances of these three tuples, the verb form 'we would have payed' is not to be found. In this case, for each tuple we look among its seen instances for identical instances (in words, POS-tags and lemmas) except for the information regarding the person, as shown in Table 1, where no useful instance has been found for T_2.

T_1 = (V[pay] , V[pagar])		
I would have payed	habría pagado	3
you would have payed	habrías pagado	1
you would have payed	pagarías	1
T_2 = (V[pay] , V[hacer] el pago)		
* would have payed	—	0
T_3 = (V[pay] it , lo V[pagar])		
I would have payed it	lo habría pagado	1

Table 1: *Seen instances in the tuples translating V[pay] that are useful to generalize 'we would have payed'.*

For each of these instances, we generate a new Spanish verb form, by changing all the information on the person in the seen form (*habría pagado*, 1stSingular) for the detected person of the expression to translate (*we*, 1stPlural). Furthermore, each new translation alternative is weighed according to the number of times the seen instance has appeared in training, shown in the last column of Table 1. This weight acts as the instance probability for these new forms. In the example, the following new forms would be generated, with probability:

T_1	we would have payed	habríamos pagado	4/6
T_1	we would have payed	pagaríamos	1/6
T_3	we would have payed it	lo habríamos pagado	1/6

Note that in the case of ambiguity (for example when generalizing a form with 'you', it can be translated into 2nd person singular or plural in Spanish), our approach is to overgenerate all possible forms and let the SMT combination of models choose the most convenient one. Actually, we expect the target Language Model to help decide the best translation alternative.

2.4. Extended generalization

In many cases, we observe only one exact realization of the test verb form in the training set. If this instance is found in a highly-improbable tuple T_i, the translation system will be forced to produce this translation, ignoring the fact that there may be several other tuples T_k translating the class with much higher probability.

Then, another approach to generalization is to look for generalization instances in all tuples, no matter whether there already is one exact seen instance of the test verb form in one tuple T_i. We will call this approach Extended Generalization. A comparison of translation results for these alternative approaches is performed in the next section.

3. Experimental results

In this section experiments translating from English into Spanish are reported. This task is especially complex in that we go from a less-inflected language with smaller vocabulary size to a highly-inflected language with bigger vocabulary size. Experiments have been carried out using the 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:

- Normalization of contracted forms for English (ie. wouldn't = would not, we've = we have)
- English POS-tagging using freely-available *TnT* tagger [7], and lemmatization using *wnmorph*, included in the WordNet package [8].
- Spanish POS-tagging using *FreeLing* analysis tool [9]. This software also generates a lemma for each input word.

3.1. Parallel corpus statistics

Table 2 shows the statistics of the data used, where each column shows number of sentences, number of words, vocabulary, and average length of a sentence, respectively.

	sent	words	vocab	avglen
Train set				
English	29998	419113	5940	14.0
Spanish		388788	9791	13.0
Dev set				
English	350	6645	841	19.0
Test set				
English	500	7412	963	14.8

Table 2: *LC-Star English-Spanish Parallel corpus statistics.*

There are 20 unseen words in the English development set (0.3% of all words), and 48 unseen words in the English test set (0.7% of all words). Three Spanish reference translations are available for both the development and the test set.

3.2. Verb forms detection and classification

We perform a knowledge-based detection of verbs using deterministic automata implementing 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, as in [10]. This unambiguous classification is done both in the English and the Spanish side, and before training. It has been previously shown that this leads to a significant improvement in the word alignment task [10].

Table 3 shows the number of detected verbs using these detection rules, and the number of different lemmas they are mapped to. For the development and test sets, the percentage of unseen verb forms and lemmas are also shown.

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, whereas Spanish tends to omit personal pronouns and contract tense information in a single inflected form.

	verbs	unseen	lemmas	unseen
Train set				
English	56419		768	
Spanish	54460		911	
Dev set				
English	856	3%	120	0%
Test set				
English	1076	5.2%	146	4.7%

Table 3: *Detected verb forms in corpus.*

3.3. Translation results

In order to evaluate the proposed classification scheme, we have integrated it into an SMT system [11] implementing a log-linear combination of:

- a tuple-based translation model $Pr(\tilde{e}_i|\tilde{f}_k)$, as done in [6]
- a target language model $Pr(e)$ as a standard Ngram LM using SRILM [12]
- a word penalty to encourage long target sentences

Four translation experiments have been carried out, whose results are shown in Table 4. On the one hand, a baseline experiment without verb forms classification (**baseline**). Secondly, an experiment with the classification but without dealing with unseen verb forms, which are not translated (**verb class**). Later on, the same experiment including the generalization of unseen verb forms described in Section 2.3 (**verb class + gen**). Finally, a last experiment also generalizing regardless of the form appearing in the training data, as discussed in Section 2.4 (**verb class + genEX**) and shown in the last row of the table. For all four experiments, the weights of each model have been optimized according to the BLEU score in the development set.

	dev set		test set	
	WER	BLEU	WER	BLEU
baseline	21.32	0.698	23.16	0.671
verb class	19.37	0.728	22.22	0.686
verb class + gen	19.27	0.727	21.65	0.692
verb class + genEX	19.25	0.729	21.62	0.689

Table 4: *English to Spanish translation results.*

3.4. Discussion

As it can be seen, the classification produces a significant improvement both in WER and BLEU, even when not dealing with unseen verb forms (around 60 verb forms in the test set). When generalizing unseen forms we achieve a further boost in performance. Note that this could hardly be achieved by a strictly statistical model, since the form to be translated is not present in the training data. Finally, even though the idea of generalizing tuples when the verb form is seen too does not harm the performance, it does not seem to provide any significant improvement either, leading to a practically identical output.

The different behavior between development and test sets can be explained in terms of the percentage of verb forms that are unseen (which is higher in the test, as shown in Table 3), leading to a bigger improvement when performing generalization. On the other hand, in the test set we have 4.7% of the

source:	I WAS TOLD that the service IS very good
baseline:	yo estaba dicho que el servicio está muy bien
verb class:	me habían dicho que el servicio está muy bien
source:	In two days' time , if YOU HAVE NOT CALLED me I WILL CANCEL the reservation
baseline:	pasado mañana fuera tiempo , si no hemos llamado anular la reserva
verb class:	en dos días tiempo , si UNSEEN UNSEEN la reserva
verb class+gen:	en dos días tiempo , si no ha llamado la anularé la reserva

Figure 1: *Examples of translated sentences. English detected verb forms are shown in capital letters.*

lemmas which are unseen and therefore cannot be translated at all unless a dictionary is provided. This effect is not present in the development set, which indicates that there is room for improvement in the final results.

Examples of translated sentences can be found in Figure 1.

4. Conclusion and future work

A linguistic classification of translation phrases, and specifically verb forms, in order to improve statistical machine translation performance has been presented. This classification allows for a better translation modeling and a generalization to unseen forms. Results in an English to Spanish task have been presented, showing significant improvements in WER and BLEU.

Yet a careful study of the output sentences shows room for further improvement. Apart from possibly dealing with unseen lemmas by means of a dictionary, we observe a certain loss of contextual information of the verb form given the previous target words. Our current instance model relies strongly on the target language model to choose the most adequate target form. Alternative methods to estimate this model should also explored, adding more accurate context features, at least for the most frequent verb lemmas. Another problem that needs to be solved is the case of Spanish enclitic pronouns, which are currently ignored by the verb forms detector due to the tagging software used. As the corresponding English pronoun does appear as a separate word, this is detected and we have an inconsistency that leads to mistakes regarding these expressions (for example, 'tell me' is badly translated into 'decir' whereas the baseline system outputs 'dígame').

As future work, experiments with a much bigger parallel corpus (European Parliament) will be carried out, evaluating this performance improvement for different sizes of the training data. Informal observation of baseline translations for this task indicate that many translation errors involve verb forms even with millions of training words.

Additionally, the next step will be to perform an straightforward classification of all simple noun phrases to the lemma of the noun. And finally, inflected adjectives in Spanish should also be tackled as a class.

5. References

[1] F. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev, "Syntax for statistical machine translation," Johns Hopkins University, Baltimore, USA, Tech. Rep. Summer Workshop Final Report, 2003.

[2] N. Ueffing and H. Ney, "Using pos information for smt into morphologically rich languages," *10th Conf. of the European Chapter of the Association for Computational Linguistics*, pp. 347–354, April 2003.

[3] M. Popovic and H. Ney, "Towards the use of word stems and suffixes for statistical machine translation," *4th Int. Conf. on Language Resources and Evaluation, LREC'04*, pp. 1585–1588, May 2004.

[4] F. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics*, vol. 30, no. 4, pp. 417–449, December 2004.

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

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

[7] T. Brants, "TnT – a statistical part-of-speech tagger," in *Proc. of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA, 2000. [Online]. Available: http://www.coli.uni-sb.de/~thorsten/tnt

[8] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, and R. Tengi, "Five papers on wordnet," *Special Issue of International Journal of Lexicography*, vol. 3, no. 4, pp. 235–312, 1991.

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

[10] A. de Gispert, "Phrase linguistic classification for improving statistical machine translation," *Accepted for Publication at the ACL 2005 Students Workshop*, June 2005.

[11] J. Crego, J. Mariño, and A. de Gispert, "An Ngram-based statistical machine translation decoder," *Submitted to Interspeech 2005*, April 2005.

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