Towards the Use of Word Stems and Suffixes for Statistical Machine Translation

Maja Popović, Hermann Ney

Lehrstuhl für Informatik VI - Computer Science Department
RWTH Aachen University
Ahornstrasse 55, 52056 Aachen, Germany
{popovic, ney}@cs.rwth-aachen.de

Abstract

In this paper we present methods for improving the quality of translation from an inflected language into English by making use of part-of-speech tags and word stems and suffixes in the source language. Results for translations from Spanish and Catalan into English are presented on the LC-STAR trilingual corpus which consists of spontaneously spoken dialogues in the domain of travelling and appointment scheduling. Results for translation from Serbian into English are presented on the Assimil language course, the bilingual corpus from unrestricted domain. We achieve up to 5% relative reduction of error rates for Spanish and Catalan and about 8% for Serbian.

1. Introduction

The goal of statistical machine translation (SMT) is to translate an input word sequence $s_1, \ldots, s_J$ in the source language into a target language word sequence $t_1, \ldots, t_I$. Given the source language sequence, we have to choose the target language sequence that maximises the product of the language model probability $Pr(t_1^I)$ and the translation model probability $Pr(s_1^J|t_1^I)$. Those two probabilities can be modelled independently of each other. The translation model describes the correspondence between the words in the source and the target sequence whereas the language model describes well-formedness of a produced target sequence. For descriptions of SMT systems see for example (Brown et al., 1993; Vogel et al., 2000).

In order to improve the translation process, it is possible to perform preprocessing steps in both source and target language sequence and, if necessary, the inverse transformations are applied to the generated output sequence. In the work presented here, we apply transformations only on the source language sequence which has the more inflective morphology. We investigate possibilities for improving the quality of translation from morphologically rich languages into English using word stems and suffixes and different types of language resources in the source language.

Source languages in our experiments are Spanish, Catalan and Serbian. Additional language resources used in the experiments are Spanish and Catalan part-of-speech (POS) tags and base forms whereas for Serbian no additional language resources were available.

2. Transformations in the Inflected Language

One of the main problems when translating an inflected language into English is a low coverage of the probabilistic lexicon: since existing SMT systems usually regard only full forms of the words, translation of full forms which have not been seen in the training corpus is not possible even if the base form or stem of the word has been seen.

Another problem is that an English word might correspond to only a part of a word in the another language. For example, the Spanish word "estamos" corresponds to the two English words "we are" (the stem "esta" corresponds to the word "are" and the suffix "mos" to the word "we").

We propose two methods to overcome these problems:

- making use of word stems and suffixes which are automatically determined
- using base forms and part-of-speech (POS) tags which are given as additional language resources

2.1. Related Work

There are many publications about discovering of word morphemes and splitting words (e.g. (Goldsmith, 2001), (Creutz and Lagus, 2002)), but the use of morphemes for statistical machine translation has been not investigated yet.

Most of the work about the use of morpho-syntactic information for SMT considers the translation from German into English. In (Nießen and Ney, 2001) hierarchical lexicon models containing base forms and POS tags are proposed. (Koehn and Knight, 2003) investigated different empirical methods for splitting German compounds.

In this work, we investigate separation of different information sources contained in one single source word using different language resources for the translation from Spanish, Catalan and Serbian into English.

2.2. Treatment of Spanish and Catalan Verbs

Spanish and Catalan have an especially rich morphology for verbs. Person and tense are expressed by the suffix so that many different full forms of one verb exist. Also, in both source languages the subject pronoun (e.g. I, we, it) is usually omitted, which often causes missing pronouns in the English translation.

We introduce two types of transformations to the verbs using two types of language resources provided by Universitat Politècnica de Catalunya (UPC) (Arranz et al., 2003) such that a verb form that is more convenient for translation into English is obtained, and the number of unseen word forms is reduced:

- **Base-POS representation**

  Additional language resources used for this method are POS tags and base forms. The full form of the verb is replaced with its base form and the sequence of relevant POS tags. POS tags corresponding to person and future or conditional tense are considered to be relevant for translation into English.

- **Stem-suffix representation**

  Additional language resources needed for this method are POS tags. A list of suffixes which corresponds to the set of relevant POS tags is defined and those suffixes are split from the stem. The suffix list is obtained by counting co-occurences of the POS tag t_s of the source word s and possible suffixes x_s: $C(t_s, x_s) = \sum_{n=1}^{N} \delta(t_{s_n}, t_s) \cdot \delta(x_{s_n}, x_s)$ where N is total number of running words. Too long and too short suffixes are dropped using a simple heuristic: if suffix x'_s is extension of the suffix x_s (for example $t_s = 1P$, one suffix is $x'_s = mos$ and the other is $x_s = os$), drop the shorter one if $\frac{C(t_s, x'_s)}{C(t_s, x_s)} > f_0$, otherwise drop the longer one. Threshold f_0 is empirically set to 0.9. Examples of transformed Spanish verbs are shown in Table 1.

original	base-POS	stem-suffix	English
estoy	estar 1S	est_ +oy	I am
estaremos	estar F 1P	esta_ +remos	we will be
tendrían	tener C 3P	tend_ +rían	they would have

Table 1: Examples of transformed Spanish verbs

2.3. Treatment of Serbian Words

Serbian as a Slavic language has a very rich morphology for all open word classes, whereby the information contained in the suffix is usually not relevant for translation into English. Therefore we reduce the words of this language into stems. We split the word into stem and suffix and then drop the suffix. Since no additional language resources were available for this language, for each word (independently of the word class) an optimal splitting point is found automatically by iterative application of the slightly modified frequency method described in (Koehn and Knight, 2003). We use harmonic mean as a metric instead of geometric mean because geometric mean always prefers splits in which either the stem or (more often) the suffix consists of a single letter. In the first iteration, counts of all possible stems s_s and suffixes x_s are collected by taking into account all possible splits (s_{s_k}, x_{s_k}) for each word s. Given these counts, for each word s we choose the split (s_s, s_x) with the highest harmonic mean of its stem and suffix count: $(s_s, s_x) = \arg\max_{(s_{s_k}, x_{s_k})} \frac{2C(s_{s_k})C(x_{s_k})}{C(s_{s_k})+C(x_{s_k})}$. If the harmonic mean of this optimal split is larger than the count of the word itself $C(s)$ the word is replaced with stem and suffix, otherwise the word is left unsplit. The new suffix and stem counts are collected from the split words, and the procedure is repeated until the possible splits do not change anymore.

Example of transformation of an adjective is presented in Table 2 (suffix depends on the gender and on the case).

original	stem	English
mali	mal_	small (boy)
mala	mal_	small (girl)
malim	mal_	(with a) small (boy)
malom	mal_	(with a) small (girl)

Table 2: Examples of reduced Serbian words

3. Experiments and Results

As already pointed out, transformations were applied in the source language, and then training and search were performed using the transformed source language data. The translation system we used is the Alignment Templates system with scaling factors (Och and Ney, 2002). Modifications of the training and search procedure were not necessary.

Evaluation metrics used in our experiments are WER (Word Error Rate), PER (Position-independent word Error Rate) and BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002). Since BLEU is an accuracy measure, we use 1-BLEU as error measure.

3.1. Translation Results for LC-STAR

The experiments are performed on the trilingual corpus which is successively built in the framework of the LC-STAR project. It contains Spanish, Catalan and English full form text, POS tags, as well as Spanish and Catalan base forms. At the time of our experiments it consisted of about 13k sentences and 120k running words.

As Table 3 shows, both preprocessing methods reduced the vocabulary size and the number of singletons in the Spanish and Catalan training corpus, as well as number of out-of-vocabulary (OOV) words in the development and test corpus. Nevertheless, they are still larger than in English corpus.

Table 4 presents an assessment of translation quality for both language pairs Catalan-English and Spanish-English. We see that there is a small but consistent decrease in all error rates for both transformation methods and for all test sets except for Spanish test corpus. A reason for this exception might be the large number of sentences containing the expression “que te parece” which corresponds to “what do you think about” in the reference English sentences whereas for the transformed source corpus this expression tends to be translated as “how about”. Therefore we performed an additional control experiment on the Spanish test set, namely additional rescoring of the best hypotheses using both the baseline system and the new system (the details of the method are described in (Och and Ney, 2001)). The results presented in Table 5 show the decrease in word error rates for both preprocessing methods. For the other test sets, rescoring does not yield significant improvements.

From translation examples (Table 6, Table 7) it can be seen that for the transformed corpus the system is able to produce correct or approximatively correct translations even if the full form has not been seen in the training corpus (marked by ”UNKNOWN_” in the baseline result example). Furthermore, the improved system is better capable of producing the correct English pronouns.

		Spanish		Catalan		English
		Original	Transformed	Original	Transformed	Original
Train	Sentences	13352				
	Words+Punctuation	118534	135316	118137	135515	123454
	Vocabulary	3933	2969	3572	2844	2154
	Singletons	1844	1314	1658	1262	790
Dev	Sentences	272				
	Words+Punctuation	2217	2531	2211	2541	2267
	OOV	35 (1.2%)	27 (1.0%)	36 (1.6%)	26 (1.0%)	21 (0.9%)
Test	Sentences	262				
	Words+Punctuation	2451	2800	2470	2819	2626
	OOV	30 (1.2%)	23 (0.8%)	35 (1.4%)	25 (0.8%)	18 (0.7%)

Table 3: Statistics of the training, develop and test set of the English-Spanish-Catalan LC-STAR corpus

		Develop			Test		
		WER	PER	1-BLEU	WER	PER	1-BLEU
Catalan	Baseline	28.3	24.1	50.1	25.4	22.3	44.5
	Base+POS	**27.7**	23.0	48.9	24.9	21.4	**42.8**
	Stem+Suffix	**27.7**	**22.8**	**48.7**	**24.6**	**21.1**	43.1
Spanish	Baseline	27.4	23.5	49.4	**24.0**	**20.8**	**42.5**
	Base+POS	27.0	22.9	48.5	25.8	21.6	44.3
	Stem+Suffix	**26.6**	**22.2**	**48.1**	24.4	20.9	43.1

Table 4: Translation error rates [%] for Catalan–English and for Spanish–English

		Test		
		WER	PER	1-BLEU
Spanish	Baseline	24.0	20.8	42.5
	+ Base+POS	23.8	20.5	**41.8**
	+ Stem+Suffix	**23.5**	**20.2**	**41.8**

Table 5: Translation error rates [%] for the rescored Spanish–English test set

3.2. Translation Results for Assimil

The experiments are performed on the small bilingual corpus containing about 3k sentences and 25k running words from unrestricted domain of the Assimil language course.

Table 8 shows the significant reduction of vocabulary size and number of singletons in the training corpus, as well as of the number of OOV words in develop and test corpus. As we can see in Table 9, there is a significant decrease in all error rates when reduction to the word stem is applied. Since the redundant information contained in the suffix is removed, the system can better capture the relevant information and is capable of producing correct translations for unseen word forms.

4. Conclusion and Future Work

In this work, we presented methods for separating different information sources contained in one single word and making use of word morphemes for statistical machine translation from inflected languages into English. Experiments showed that the use of word morphemes improves the translation quality.

		Serbian		English
		Original	Transformed	Original
Train	Sent.	2926		
	Voc.	4923	3712	2898
	Singl.	2988	1998	1370
Dev	Sent.	100		
	OOV	63 (9.0%)	39 (5.6%)	21 (2.6%)
Test	Sent.	100		
	OOV	153 (15.6%)	107 (10.9%)	82 (7.6%)

Table 8: Statistics of the training, develop and test set of the English-Serbian Assimil corpus

	Develop			Test		
	WER	PER	1-BLEU	WER	PER	1-BLEU
Baseline	40.9	36.1	69.1	51.2	44.3	79.6
Stem	**37.5**	**33.5**	**63.8**	**48.3**	**42.4**	**75.7**

Table 9: Translation error rates [%] for Serbian–English

For our best translation system that was taken as baseline, we achieve up to 5% relative reduction of error rates for Spanish and Catalan and about 8% for Serbian by applying transformations.

Additional language resources we used were base forms and POS tags for Spanish and Catalan. The stem-suffix representation of Spanish and Catalan verbs has yielded similar translation quality as the base-POS representation, and the advantage of this method is that it requires less complex language resources, i.e. only POS tags.

d'acord , i *treballarem* .	⇒	d'acord , i *treballar F 1P* .
	verb transformations	d'acord , i *treballa_ +rem* .
⇓ C → E (baseline)		⇓ C → E
okay , and *UNKNOWN_treballarem* .		okay , and *we will work* .
sí *tornarem* dimecres al vespre .	⇒	sí *tornar F 1P* dimecres al vespre .
	verb transformations	sí *torna_ +rem* dimecres al vespre .
⇓ C → E (baseline)		⇓ C → E
yes fly back on Wednesday evening .		yes *we will* fly back on Wednesday evening .

Table 6: Examples of Catalan–English translations with and without verb transformations

de acuerdo , y *trabajaremos* .	⇒	de acuerdo , y *trabajar F 1P* .
	verb transformations	de acuerdo , y *trabaja_ +remos* .
⇓ S → E (baseline)		⇓ S → E
okay , and *UNKNOWN_trabajaremos* .		okay , and *we will work* .
creo que *cogeré* el tren , ahí .	⇒	creo que *coger F 1S* el tren , ahí .
	verb transformations	creo que *coge_ +ré* el tren , ahí .
⇓ S → E (baseline)		⇓ S → E
I think take the train , there .		I think *I will* take the train , there .

Table 7: Examples of Spanish–English translations with and without verb transformations

Reducing Serbian words into stems has significantly reduced the translation errors on the small corpus even without using additional language resources.

We plan to apply these methods to other tasks and other language pairs. We will also investigate possibilities for improving the quality of the other translation direction (from English into the highly inflected language) using morpho-syntactic knowledge.

5. Acknowledgements

This work was partly supported by the LC-STAR project by the European Community (IST project ref. no. 2001- 32216).

6. References

V. Arranz, N. Castell, and J. Giménez. 2003. Development of language resources for speech-to-speech translation. In *Proc. of RANLP'03*, Borovets, Bulgaria, September.

P. F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311

M. Creutz and K. Lagus. 2002. Unsupervised discovery of morphemes. In *Proc. of the Workshop on Morphological and Phonological Learning of ACL-02)*, pages 21–30, Philadelphia, PA, July.

J. Goldsmith. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2):153–198.

P. Koehn and K. Knight. 2003. Empirical Methods for Compound Splitting. In *Proc. of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 347–354, Budapest, Hungary, April.

S. Nießen and H. Ney. 2001. Toward hierarchical models for statistical machine translation of inflected languages. In *39th Annual Meeting of the Assoc. for Computational Linguistics - joint with EACL 2001: Proc. Workshop on Data-Driven Machine Translation*, pages 47–54, Toulouse, France, July.

F. J. Och and H. Ney. 2001. Statistical multi-source Translation. In *Proc. of Machine Translation Summit VIII*, pages 253–258, Santiago de Compostela, Galicia, Spain, September.

F. J. Och and H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, July.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Assoc. for Computational Linguistics*, pages 311–318, Philadelphia, PA, July.

S. Vogel, F.J. Och, C. Tillmann, S. Nießen, H. Sawaf, and H. Ney. 2000. Statistical methods for machine translation. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 377–393. Springer Verlag: Berlin, Heidelberg, New York.