Statistical Machine Translation of Serbian-English

Maja Popović (1), Slobodan Jovičić, Zoran Šarić (2)

(1) Lehrstuhl für Informatik VI - Computer Science Department,
RWTH Aachen University, Ahornstrasse 55, 52056 Aachen, Germany
popovic@informatik.rwth-aachen.de
(2) School of Electrical Engineering, Bulevar kralja Aleksandra 73,
11000 Beograd, Serbia and Montenegro
jovicic@etf.bg.ac.yu

Abstract

In this work we present the first results of statistical approach to the machine translation of Serbian language into English and vice versa. The experiments are performed on the Assimil language course, bilingual parallel corpus which consists of about 3k sentences and 20k running words from unrestricted domain. The error rates for the translation of Serbian into English are about 35-45% and for the other direction about 45-55%. The results are comparable with those for the other language pairs having been translated using statistical approach. Reducing Serbian words into stems has decreased error rates for the translation into English for about 8% relative.

1. Introduction

Statistical approach to machine translation has been receiving more and more attention. The goal of statistical machine translation is to translate a source language sequence $f_1, \ldots, f_J$ into a target language sequence $e_1, \ldots, e_I$ by maximising the conditional probability $Pr(e_1^I|f_1^J)$. This approach has been applied on various languages (e.g. English, French, German, Spanish, Chinese, Japanese, etc.) and on different domains (touristical information, travelling and appointment schedulling, parliamentar debates) and has shown to obtain very good results in comparison to other classical approaches. However, for the Serbian language this approach has not been tested so far.

In this work, we present the first results for Serbian-English language pair. Translation experiments have been done on the relatively small bilingual corpus from unrestricted domain. The baseline experiments have been done in both translation directions, and translation from Serbian into English has been additionaly improved by reducing Serbian words into stems.

2. Statistical Machine Translation

The main concept of statistical machine translation (SMT) is to translate a source word sequence $f_1, \ldots, f_J$ into a target language word sequence $e_1, \ldots, e_I$ using probability models.

Given the source language sequence f_1^J, we have to choose the target language sequence $\hat{e}_1^I$ that maximises the probability $Pr(e_1^I|f_1^J)$:

$$\hat{e}_1^I = \arg\max_{e_1^I} Pr(e_1^I|f_1^J)$$

This probability can be represented as a product of the language model probability $Pr(e_1^I)$ and the translation model probability $Pr(f_1^J|e_1^I)$:

$$Pr(e_1^I|f_1^J) \quad = \quad Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)$$

Those two probabilities can be modelled independently of each other.

The translation model describes the correspondence between the words in the source sequence and the words in the target sequence whereas the language model describes well-formedness of a produced target sequence.

The translation model can be rewritten in the following way:

$$Pr(f_1^J|e_1^I) \quad = \quad \sum_{a_1^J} Pr(f_1^J, a_1^J|e_1^I)$$

where a_1^J are called alignments and represent a mapping from the source word position j to the target word position $i = a_j$. Alignments are introduced into translation model as a hidden variable, similar to the concept of Hidden Markov Models (HMM) in speech recognition.

The translation probability $Pr(f_1^J, a_1^J|e_1^I)$ can be further rewritten as a product over the words in

the source sentence and decomposed as follows:

$$\begin{aligned} Pr(f_1^J, a_1^J|e_1^I) &= \prod_{j=1}^{J} Pr(f_j, a_j|f_1^{j-1}, a_1^{j-1}, e_1^I) \\ &= \prod_{j=1}^{J} Pr(a_j|f_1^{j-1}, a_1^{j-1}, e_1^I) \cdot \\ &\quad \cdot Pr(f_j|f_1^{j-1}, a_1^j, e_1^I) \end{aligned}$$

where $Pr(a_j|f_1^{j-1}, a_1^{j-1}, e_1^I)$ is called alignment probability and $Pr(f_j|f_1^{j-1}, a_1^j, e_1^I)$ is lexicon probability.

These probability models replace true but unknown probability distributions. The parameters of the models (i.e. the probabilities) have to be learnt from the parallel bilingual text using suitable training criterion. Traditionally, the so-called maximum likelihood criterion is used.

The generation of the target sentence takes into account all three knowledge sources: alignment model, lexicon model and language model. An efficient implementation of finding a sequence which corresponds to the maximum probability has to be found. The details of the generation process very much depend on the specific structure of used probability models.

For detailed descriptions of SMT systems and models see for example [1], [2],[3],[4].

3. Experiments and Results

3.1. Corpus

The corpus used in this work is the small bilingual corpus of the Assimil language course containing about 3k sentences and 25k running words. Since the domain of the corpus is not restricted, the vocabulary size (Voc) and the number of singletons (Singl) in the training corpus is rather large, as well as the number of out of vocabulary words (OOV) in the development and test corpus. Due to the very rich inflectional morphology of the Serbian language, all those numbers are much larger for this language than for English. By reducing Serbian words into stems (transformed Serbian) these numbers are decreased, but they still remain above the values for English language.

Detailed corpus statistics is shown in Table 1.

Table 1: Statistics of the training, develop and test set of the English-Serbian Assimil corpus

		Serbian		English
		Original	Transformed	Original
Train	Sent	2926		
	Words	24725	24725	27471
	Voc	4923	3712	2898
	Singl	2988	1998	1370
Dev	Sent	100		
	Words	696	696	790
	OOV	9.0%	5.6%	2.6%
Test	Sent	100		
	Words	980	980	1083
	OOV	15.6%	10.9%	7.6%

3.2. Experiments

The translation experiments are performed in both directions, i.e. from Serbian into English and other way round. For the translation into English, two types of Serbian corpora have been used: original and transformed - with the words reduced into stems.

Since 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, we also applied reduction of the words of this language into stems [5].

The word is first splitted into stem and suffix and then the suffix is dropped. Since POS tags or similar additional information were not available, an optimal splitting point for each word is found automatically by iterative application of the slightly modified frequency method described in [6]. This method has been proposed for splitting German compound words. The compound is broken into its components if the geometric mean of the component counts (frequencies) is larger than the count (frequency) of the compound itself. In our experiments 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 we calculate harmonic mean for all possible splits:

$$HM(s_{s_k}, s_{x_k}) = \frac{2 \cdot C(s_{s_k}) \cdot C(x_{s_k})}{C(s_{s_k}) + C(x_{s_k})}$$

and choose the split (s_s, s_x) with the highest harmonic mean as optimal:

$$(s_s, s_x) = \arg \max_{(s_{s_k}, x_{s_k})} HM(s_{s_k}, s_{x_k})$$

If the count (frequency) of the word itself $C(s)$ is larger than the harmonic mean of its optimal split, the word is left unsplit, otherwise is replaced with the stem and the suffix of the optimal split.

In the next iteration, the new suffix and stem counts are collected from the new text taking into account 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).

Table 2: Examples of reduced Serbian words

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

The translation system we used is the Alignment Templates system with scaling factors [7]. Modifications of the training and search procedure were not necessary for the translation of the transformed Serbian corpus.

3.3. Translation Results

Evaluation metrics used in our experiments are WER (Word Error Rate), PER (Position-independent word Error Rate) and BLEU (BiLingual Evaluation Understudy) [8]. Since BLEU is an accuracy measure, we use 1-BLEU as error measure.

Error rates for the translation from Serbian into English are shown in Table 3 and some translation examples can be seen in Table 5. It can be seen that 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 or approximatively correct translations even for unseen full forms of the words (marked by "UNKNOWN_" in the baseline result example).

Table 4 shows results for the translation from English into Serbian. As expected, all error rates

Table 3: Translation error rates [%] for Serbian→English

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

Table 4: Translation error rates [%] for English→Serbian

En→Sr	Develop		
	WER	PER	1-BLEU
Baseline	46.1	41.0	76.5
	Test		
	WER	PER	1-BLEU
Baseline	55.3	48.7	80.3

are higher than for the other translation direction since translation into the morphologically richer language is always more difficult. In Table 6 we can see two examples of wrong Serbian full form words. In the first sentence there are three words which are translated into the wrong case and/or gender, but the sentence still conveys the correct semantics. On the contrary, in the other sentence the wrong form of the verb induces semantical error because Serbian is the pro-drop language (the pronoun is often omitted and the information about the person as well as the tense is contained in the suffix). The obtained translation indicates that he was asking for extension 35 (third person singular, past tense), but the correct meaning is that you should ask for extension 35 (second person plural, imperative mood).

The error rates for both translation directions are comparable with those for other language pairs especially when the facts that the corpus is rather small, domain is unrestricted, and morphology of Serbian language is very rich are taken into account.

We believe that the morpho-syntactic analysis of the Serbian language can improve the results further (like for example in [9],[10]).

Table 5: Examples of Serbian–English translations with and without transformations

to je mali grip , ništa *ozbiljno* .	⇒ transformations	to je mal_ grip , ništa *ozbilj_* .
⇓ Sr → En (baseline)		⇓ Sr' → En
it is a touch of flu , nothing *UNKNOWN_ozbiljno* .		it is a small flu , nothing *serious*.
hajde da pogledamo neki izraz sa *glagolom* ''get'' .	⇒ transformations	hajde da pogleda_ nek_ izraz_ sa *glagol_* ''get'' .
⇓ Sr → En (baseline)		⇓ Sr' → En
let us look at some expressions with *UNKNOWN_glagolom* ''get'' .		let us look at some expressions with *verbs* ''get''
svi su u isto vreme *pokušavali* da udju u autobus .	⇒ transformations	svi su u ist_ vreme *pokušav_* da udj_ u autobus .
⇓ Sr → En (baseline)		⇓ Sr' → En
everyone in as *UNKNOWN_pokušavali* time to in in in bus .		everyone in same time *trying* to come at the bus .

Table 6: Examples of English–Serbian translations

here is Peter and his friend Anne .	
⇓ En → Sr (baseline)	reference translation:
evo *Peter* i *njegov prijatelj* Anne .	evo *Petera* i *njegove prijateljice* Anne .
ask for extension thirty five .	
⇓ En → Sr (baseline)	reference translation:
tražio lokal trideset pet .	*tražite* lokal trideset pet .

4. Conclusions

In this work, we presented the results for the statistical approach to the machine translation of language pair Serbian-English. Obtained results are comparable with those for the other language pairs treated by statistical methods. With, to our knowledge, the best SMT system we obtained the error rates of about 45% for the translation into English and about 50% for the other direction. Reduction of words into stems has improved the translation into English for about 8% relative in comparison to the baseline system.

We believe that the results can be further improved by the use of different kind of morpho-syntactic knowledge (e.g. POS tags, base forms, etc.). We will also investigate possibilities for improvement of the other translation direction.

5. References

[1] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[2] S. Vogel, F. J. Och, C. Tillmann, S. Nießen, H. Sawaf, and H. Ney, "Statistical methods for machine translation," in *Verbmobil: Foundations of Speech-to-Speech Translation* (W. Wahlster, ed.), pp. 377–393, Springer Verlag: Berlin, Heidelberg, New York, 2000.

[3] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, pp. 19–51, March 2003.

[4] S. Vogel, H. Ney, and C. Tillmann, "HMM-based word alignment in statistical translation," in *Proc. 16th Int. Conf. on Computational Linguistics (COLING)*, (Copenhagen, Denmark), pp. 836–841, Aug. 1996.

[5] M. Popović and H. Ney, "Towards the use of word stems and suffixes for statistical machine translation," in *Proc. 4th Int. Conf. on Language Resources and Evaluation (LREC)*, (Lisbon, Portugal), pp. 1585–1588, May 2004.

[6] P. Koehn and K. Knight, "Empirical methods for compound splitting," in *Proc. 10th Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL)*, (Budapest, Hungary), pp. 347–354, April 2003.

[7] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proc. 40th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, (Philadelphia, PA), pp. 295–302, July 2002.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. 40th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, (Philadelphia, PA), pp. 311–318, July 2002.

[9] S. Nießen and H. Ney, "Morpho-syntactic analysis for reordering in statistical machine translation," in *Proc. MT Summit VIII*, (Santiago de Compostela, Galicia, Spain), pp. 247–252, September 2001.

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