Augmenting a Small Parallel Text with Morpho-syntactic Language Resources for Serbian-English Statistical Machine Translation

Maja Popović, David Vilar, Hermann Ney
Lehrstuhl für Informatik VI
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
{popovic,vilar,ney}@informatik.rwth-aachen.de

Slobodan Jovičić, Zoran Šarić
Faculty of Electrical Engineering
University of Belgrade
Serbia and Montenegro
jovicic@etf.bg.ac.yu

Abstract

In this work, we examine the quality of several statistical machine translation systems constructed on a small amount of parallel Serbian-English text. The main bilingual parallel corpus consists of about 3k sentences and 20k running words from an unrestricted domain. The translation systems are built on the full corpus as well as on a reduced corpus containing only 200 parallel sentences. A small set of about 350 short phrases from the web is used as additional bilingual knowledge. In addition, we investigate the use of monolingual morpho-syntactic knowledge i.e. base forms and POS tags.

1 Introduction and Related Work

The goal of statistical machine translation (SMT) 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 probability can be factorised into the translation model probability $P(f_1^J|e_1^I)$ which describes the correspondence between the words in the source and the target sequence, and the language model probability $P(e_1^I)$ which describes well-formedness of the produced target sequence. These two probabilities can be modelled independently of each other. For detailed descriptions of SMT models see for example (Brown et al., 1993; Och and Ney, 2003).

Translation probabilities are learnt from a bilingual parallel text corpus and language model probabilities are learnt from a monolingual text in the target language. Usually, the performance of a translation system strongly depends on the size of the available training corpus. However, acquisition of a large high-quality bilingual parallel text for the desired domain and language pair requires lot of time and effort, and, for many language pairs, is even not possible. Besides, small corpora have certain advantages - the acquisition does not require too much effort and also manual creation and correction are possible. Therefore there is an increasing number of publications dealing with limited amounts of bilingual data (Al-Onaizan et al., 2000; Nießen and Ney, 2004).

For the Serbian language, as a rather minor and not widely studied language, there are not many language resources available, especially not parallel texts. On the other side, investigations on this language may be quite useful since the majority of principles can be extended to the wider group of Slavic languages (e.g. Czech, Polish, Russian, etc.).

In this work, we exploit small Serbian-English parallel texts as a bilingual knowledge source for statistical machine translation. In addition, we investigate the possibilities for improving the translation quality using morpho-syntactic information in the source language. Some preliminary translation results on this language pair have been reported in (Popović et al., 2004; Popović and Ney, 2004), but no systematic investigation has been done so far. This work presents several translation systems created with different amounts and types of training data and gives a detailed description of the language resources used.

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 41–48,
Ann Arbor, June 2005.

2 Language Resources

2.1 Language Characteristics

Serbian, as a Slavic language, has a very rich inflectional morphology for all open word classes. There are six distinct cases affecting not only common nouns but also proper nouns as well as pronouns, adjectives and some numbers. Some nouns and adjectives have two distinct plural forms depending on the number (if it is larger than four or not). There are also three genders for the nouns, pronouns, adjectives and some numbers leading to differences between the cases and also between the verb participles for past tense and passive voice.

As for verbs, person and many tenses are expressed by the suffix, and the subject pronoun (e.g. I, we, it) is often omitted (similarly as in Spanish and Italian). In addition, negation of three quite important verbs, "biti" (to be, auxiliary verb for past tense, conditional and passive voice), "imati" (to have) and "hteti" (to want, auxiliary verb for the future tense), is done by adding the negative particle to the verb as a prefix.

As for syntax, Serbian has a quite free word order, and there are no articles, neither indefinite nor definite.

All these characteristics indicate that morpho-syntactic knowledge might be very useful for statistical machine translation involving Serbian language, especially when only scarce amounts of parallel text are available.

2.2 Parallel Corpora

Finding high-quality bilingual or multilingual parallel corpora involving Serbian language is a difficult task. For example, there are several web-sites with the news in both Serbian and English (some of them in other languages as well), but these texts are only comparable and not parallel at all. To our knowledge, the only currently available Serbian-English parallel text suitable for statistical machine translation is a manually created electronic version of the Assimil language course which has been used for some preliminary experiments in (Popović et al., 2004; Popović and Ney, 2004). We have used this corpus for systematical investigations described in this work.

2.2.1 Assimil Language Course

The electronic form of Assimil language course contains about 3k sentences and 25k running words of various types of conversations and descriptions as well as a few short newspaper articles. Detailed corpus statistics can be seen in Table 1. Since the domain of the corpus is basically not restricted, the vocabulary size is relatively large. Due to the rich morphology, the vocabulary for Serbian is almost two times larger than for English. The average sentence length for Serbian is about 8.5 words per sentence, and for English about 9.5. This difference is mainly caused by the lack of articles and omission of some subject pronouns in Serbian .

The development and test set (500 sentences) are randomly extracted from the original corpus and the rest is used for training (referred to as 2.6k).

In order to investigate the scenario with extremely scarce training material, a reduced training corpus (referred to as 200) has been created by random extraction of 200 sentences from the original training corpus.

The morpho-syntactic annotation of the English part of the corpus has been done by the constraint grammar parser ENGCG for morphological and syntactic analysis of English language. For each word, this tool provides its base form and sequence of morpho-syntactic tags.

For the Serbian corpus, to our knowlegde there is no available tool for automatic annotation of this language. Therefore, the base forms have been introduced manually and the POS tags have been provided partly manually and partly automatically using a statistical maximum-entropy based POS tagger similar to the one described in (Ratnaparkhi, 1996). First, the 200 sentences of the reduced training corpus have been annotated completely manually. Then the first 500 sentences of the rest of the training corpus have been tagged automatically and the errors have been manually corrected. Afterwards, the POS tagger has been trained on the extended corpus (700 sentences), the next 500 sentences of the rest are annotated, and the procedure has been repeated until the annotation has been finished for the complete corpus.

Table 1: Statistics of the Serbian-English Assimil corpus

		Serbian		English	
Training:		original	base forms	original	no article
full corpus	Sentences	2632		2632	
(2.6k)	Running Words + Punct.	22227		24808	23308
	Average Sentence Length	8.4		9.5	8.8
	Vocabulary Size	4546	2605	2645	2642
	Singletons	2728	1253	1211	
reduced corpus	Sentences	200		200	
(200)	Running Words + Punct.	1666		1878	1761
	Average Sentence Length	8.3		10.4	8.8
	Vocabulary Size	778	596	603	600
	Singletons	618	417	395	
Dev+Test	Sentences	500		500	
	Running Words + Punct.	4161		4657	4362
	Average Sentence Length	8.3		9.3	8.7
	Vocabulary Size	1457	1030	1055	1052
	Running OOVs - 2.6k	12.1%	5.2%	4.8%	
	Running OOVs - 200	34.5%	27.6%	21.4%	
	OOVs - 2.6k	32.7%	19.5%	19.7%	
	OOVs - 200	76.2%	66.0%	66.8%	
External Test	Sentences	22		22	
	Running Words + Punct.	395		446	412
	Average Sentence Length	18.0		20.3	18.7
	Vocabulary Size	213	176	202	199
	Running OOVs - 2.6k	44.3%	35.4%	32.1%	34.7%
	Running OOVs - 200	53.7%	44.6%	43.7%	47.3 %
	OOVs - 2.6k	61.5%	45.4%	44.0%	44.7%
	OOVs - 200	74.6%	63.1%	63.9%	64.8%

Table 2: Statistics of the Serbian-English short phrases

		Serbian		English	
Phrases		original	base forms	original	no article
	Entries	351	351	351	351
	Running Words + Punct.	617	617	730	700
	Average Entry Length	1.8	1.8	2.1	2.0
	Vocabulary Size	335	303	315	312
	Singletons	239	209	209	208
New Running	2.6k	20.6%	14.4%	11.8%	11.8%
Words	200	50.6%	41.3%	36.7%	37.8%
New Vocabulary	2.6k	30.1%	22.1%	21.6%	21.2%
Words	200	70.7%	63.0%	63.2%	63.1%

2.2.2 Short Phrases

The short phrases used as an additional bilingual knowledge source in our experiments have been collected from the web and contain about 350 standard words and short expressions with an average entry length of 1.8 words for Serbian and 2 words for English. Table 2 shows that about 30% of words from the phrase vocabulary are not present in the original Serbian corpus and about 70% of those words are not contained in the reduced corpus. For the English language those numbers are smaller, about 20% for the original corpus and 60% for the reduced one. These percentages are indicating that this parallel text, although very scarce, might be an useful additional training material.

The phrases have also been morpho-syntactically annotated in the same way as the main corpus.

2.2.3 External Test

In addition to the standard development and test set described in Section 2.2.1, we also tested our translation systems on a short external parallel text collected from the BBC News web-site containing 22 sentences about relations between USA and Ukraine after the revolution. As can be seen in Table 1, this text contains very large portion of out-of-vocabulary words (almost two thirds of Serbian words and almost half of English words are not seen in the training corpus), and has an average sentence length about two times larger than the training corpus.

3 Transformations in the Source Language

Standard SMT systems usually regard only full forms of the words, so that translation of full forms which have not been seen in the training corpus is not possible even if the base form has been seen. Since the inflectional morphology of the Serbian language is very rich, as described in Section 2.1, we investigate the use of the base forms instead of the full forms to overcome this problem for the translation into English. We propose two types of transformations of the Serbian corpus: conversion of the full forms into the base forms and additional treatment of the verbs.

For the other translation direction, we propose removing the articles in the English part of the corpus as the Serbian language does not have any.

3.1 Transformations of the Serbian Text

3.1.1 Base Forms

Serbian full forms of the words usually contain information which is not relevant for translation into English. Therefore, we propose conversion of all Serbian words in their base forms. Although for some other inflected languages like German and Spanish this method did not yield any translation improvement, we still considered it as promising because the number of Serbian inflections is considerably higher than in the other two languages. Table 1 shows that this transformation significantly reduces the Serbian vocabulary size so that it becomes comparable to the English one.

3.1.2 Treatment of Verbs

Inflections of Serbian verbs might contain relevant information about the person, which is especially important when the pronoun is omitted. Therefore, we apply an additional treatment of the verbs. Whereas all other word classes are still replaced only by their base forms, for each verb a part of the POS tag referring to the person is taken and the verb is converted into a sequence of this tag and its base form. For the three verbs described in Section 2.1, the separation of the negative particle is also applied: each negative full form is transformed into the sequence of the POS tag, negative particle and base form. The detailed statistics of this corpus is not reported since there are no significant changes, only the number of running words and average sentence length increase thus becoming closer to the values of the English corpus.

3.2 Transformations of the English Text

3.2.1 Removing Articles

Since the articles are one of the most frequent word classes in English, but on the other side there are no arcticles at all in Serbian, we propose removing the articles from the English corpus for translation into Serbian. Each English word which has been detected as an article by means of its POS tag has been removed from the corpus. In Table 1, it can be seen that this method significantly reduces the number of running words and the average sentence length of the English corpus thus becoming comparable to the values of the Serbian corpus.

4 Translation Experiments and Results

4.1 Experimental Settings

In order to systematically investigate the impact of the bilingual training corpus size and the effects of the morpho-syntactic information on the translation quality, the translation systems were trained on the full training corpus (2.6k) and on the reduced training corpus (200), both with and without short phrases. The translation is performed in both directions, i.e. from Serbian to English and other way round. For the Serbian to English translation systems, three versions of the Serbian corpus have been used: original (baseline), base forms only (sr_base) and base forms with additional treatment of the verbs (sr_base+v-pos). For the translation into Serbian, the systems were trained on two versions of the English corpus: original (baseline) and without articles (en_no-article).

The baseline translation system is the Alignment Templates system with scaling factors (Och and Ney, 2002). Word alignments are produced using GIZA++ toolkit without symmetrisation (Och and Ney, 2003). Preprocessing of the source data has been done before the training of the system, therefore modifications of the training and search procedure were not necessary for the translation of the transformed source language corpora.

Although the development set has been used to optimise the scaling factors, results obtained for this set do not differ from those for the test set. Therefore only the joint error rates (Development+Test) are reported.

As for the external test set, results for this text are reported only for the full corpus systems, since for the reduced corpus the error rates are higher but the effects of using phrases and morpho-syntactic information are basically the same.

4.2 Translation Results

The 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 an error measure.

4.2.1 Translation from Serbian into English

Error rates for the translation from Serbian into English are shown in Table 3 and some examples are shown in Table 6. It can be seen that there is a significant decrease in all error rates when the full forms are replaced with their base forms. Since the redundant information contained in the inflection 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). The treatment of the verbs yields some additional improvements.

From the first translation example in Table 6 it can be seen how the problem of some out-of-vocabulary words can be overcomed with the use of the base forms. The second and third example are showing the advantages of the verb treatment, the third one illustrates the effect of separating the negative particle.

Reduction of the training corpus to only 200 sentences (about 8% of the original corpus) leads to a loss of error rates of about 45% relative. However, the degradation is not higher than 35% if phrases and morpho-syntactic information are available in addition to the reduced corpus.

The use of the phrases can improve the translation quality to some extent, especially for the systems with the reduced training corpus, but these improvements are less remarkable than those obtained by replacing words with the base forms.

The best system with the complete corpus as well as the best one with the reduced corpus use the phrases and the transformed Serbian corpus where the verb treatment has been applied.

4.2.2 Translation from English into Serbian

Table 4 shows results for the translation from English into Serbian. As expected, all error rates are higher than for the other translation direction. Translation into the morphologically richer language always has poorer quality because it is difficult to find the correct inflection.

The performance of the reduced corpus is degraded for about 40% relative for the baseline system and for about 30% when the phrases are used and the transformation of the English corpus has been applied.

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

Serbian → English		Development+Test		
Training Corpus	Method	WER	PER	1-BLEU
2.6k	baseline	45.6	39.6	70.0
2.6k	sr_base	43.5	38.2	68.9
2.6k	sr_base+v-pos	42.5	35.3	66.2
2.6k+phrases	baseline	46.0	39.6	69.5
2.6k+phrases	sr_base	44.6	39.1	70.2
2.6k+phrases	sr_base+v-pos	**42.1**	**35.3**	**66.0**
200	baseline	66.5	61.1	91.6
200	sr_base	63.2	58.2	90.3
200	sr_base+v-pos	63.3	56.2	88.5
200+phrases	baseline	65.2	59.5	90.2
200+phrases	sr_base	62.3	56.9	87.7
200+phrases	sr_base+v-pos	**61.3**	**53.2**	**86.2**

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

English → Serbian		Development+Test		
Training Corpus	Method	WER	PER	1-BLEU
2.6k	baseline	53.1	46.9	78.6
2.6k	en_no-article	52.6	47.2	79.4
2.6k+phrases	baseline	52.5	**46.5**	**76.6**
2.6k+phrases	en_no-article	**52.3**	47.0	79.6
200	baseline	73.6	68.0	93.0
200	en_no-article	71.5	66.5	93.4
200+phrases	baseline	71.7	66.7	92.3
200+phrases	en_no-article	**67.9**	**62.9**	**92.1**

Table 5: Translation error rates [%] for the external test

Serbian → English		External Test		
Training Corpus	Method	WER	PER	1-BLEU
2.6k	baseline	72.2	64.8	92.2
2.6k	sr_base	**66.8**	61.4	**86.9**
2.6k	sr_base+v-pos	67.5	61.4	88.3
2.6k+phrases	baseline	71.3	63.9	91.9
2.6k+phrases	sr_base	67.0	**61.2**	88.4
2.6k+phrases	sr_base+v-pos	69.7	**61.2**	89.8
English → Serbian				
2.6k	baseline	85.3	77.0	96.4
2.6k	en_no-article	**77.5**	**69.9**	95.8
2.6k+phrases	baseline	84.1	74.9	95.2
2.6k+phrases	en_no-article	77.7	70.1	**94.8**

The importance of the phrases seems to be larger for this translation direction. Removing the English articles does not have the significant role for the translation systems with full corpus, but for the reduced corpus it has basically the same effect as the use of phrases. The best system with the reduced corpus has been built with the use of phrases and removal of the articles.

Table 7 shows some examples of the translation into Serbian with and without English articles. Although these effects are not directly obvious, it can be seen that removing of the redundant information enables better learning of the relevant information so that system is better capable of producing semantically correct output. The first example illustrates an syntactically incorrect output with the wrong inflection of the verb ("čitam" means "I read"). The output of the system without articles is still not completely correct, but the semantic is completely preserved. The second example illustrates an output produced by the baseline system which is neither syntactically nor semantically correct ("you have I drink"). The output of the new system still has an error in the verb, informal form of "you" instead of the formal one, but nevertheless both the syntax and semantics are correct.

4.2.3 Translation of the External Text

Translation results for the *external test* can be seen in Table 5. As expected, the high number of out-of-vocabulary words results in very high error rates. Certain improvement is achieved with the phrases, but the most significant improvements are yielded by the use of Serbian base forms and removal of English articles. Verb treatment in this case does not outperform the base forms system, probably because there are not so many different verb forms as in the other corpus, and only a small number of pronouns is missing.

5 Conclusions

In this work, we have examined the possibilities for building a statistical machine translation system with a small bilingual Serbian-English parallel text. Our experiments showed that the translation results for this language pair are comparable with results for other language pairs, especially if the small size of the corpus, unrestricted domain and rich inflectional morphology of Serbian language are taken into account. With the baseline system, we obtained about 45% WER for translation into English and about 53% for translation into Serbian.

We have systematically investigated the impact of the corpus size on translation quality, as well as the importance of additional bilingual knowledge in the form of short phrases. In addition, we have shown that morpho-syntactic information is a valuable language resource for translation of this language pair.

Depending on the availability of resources and tools, we plan to examine parallel texts with other languages, and also to do further investigations on this language pair. We believe that more refined use of the morpho-syntactic information can yield better results (for example the hierarchical lexicon model proposed in (Nießen and Ney, 2001)). We also believe that the use of the conventional dictionaries could improve the Serbian-English translation.

Acknowledgement

This work was partly funded by the Deutsche Forschungsgemeinschaft (DFG) under the project "Statistical Methods for Written Language Translation" (Ne572/5).

References

Y. Al-Onaizan, U. Germann, U. Hermjakob, K. Knight, P. Koehn, D. Marcu, and K. Yamada. 2000. Translating with scarce resources. In *National Conference on Artificial Intelligence (AAAI)*.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Sonja Nießen and Hermann 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.

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

Franz J. Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical

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

to je suvishe *skupo* .	⇒ base forms	to biti suvishe *skup* .	⇒ verb treatment	to SG3 biti suvishe *skup* .
⇓ Sr → En (baseline)		⇓ Sr' → En		⇓ Sr" → En
it is too *UNKNOWN_skupo* .		it is too *expensive* .		it is too *expensive* .
on ne *igra* .	⇒ base forms	on ne *igrati* .	⇒ verb treatment	on ne *SG3 igrati* .
⇓ Sr → En (baseline)		⇓ Sr' → En		⇓ Sr" → En
he he does not .		he *do not* play .		he *does not play* .
da , ali *nemam* mnogo vremena .	⇒ base forms	da , ali *nemati* mnogo vreme .	⇒ verb treatment	da , ali *SG1 ne imati* mnogo vreme .
⇓ Sr → En (baseline)		⇓ Sr' → En		⇓ Sr" → En
yes , but *I have* much time .		yes , but *not* much time .		yes , but *I have not got* much time .

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

you should not read in bed .	⇒ remove articles	you should not read in bed .	
⇓ En → Sr (baseline)		⇓ En' → Sr	reference translation:
treba ne čitam u krevet .		ne bi trebalo čitate u krevet .	ne bi trebalo da čitate u krevetu .
have a drink .	⇒ remove articles	have drink .	
⇓ En → Sr (baseline)		⇓ En' → Sr	reference translation:
imate pijem .		uzmi nešto za piće .	uzmite nešto za piće .

machine translation. In *Proc. 40th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, July.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Kishore Papineni, Salim Roukos, Todd Ward, and Wie-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.

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

M. Popović, S. Jovičić, and Z. Šarić. 2004. Statistical machine translation of Serbian-English. In *Proc. of Int. Workshop on Speech and Computer (SPECOM)*, pages 410–414, St. Petersburg, Russia, September.

A. Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proc. Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 133–142, Sommerset, NJ.