Slovenian to English Machine Translation using Corpora of Different Sizes and Morpho-syntactic Information

Mirjam Sepesy Maučec,[*] Janez Brest,[†] Zdravko Kačič[*]

[*]Institute of Electronic and Telecommunication
Faculty of Electrical Engineering and Computer Science
University of Maribor
Smetanova 17, SI-2000 Maribor
mirjam.sepesy@uni-mb.si, kacic@uni-mb.si

[†]Institute of Computer Science
Faculty of Electrical Engineering and Computer Science
Smetanova 17, SI-2000 Maribor
janez.brest@uni-mb.si

Abstract

Word based statistical machine translation has emerged as a robust method for building machine translation systems. Inflective languages point out some problems with the approach. Data sparsity is one of them. It can be partly solved by enlarging the training corpus and/or including richer linguistic information: lemmas and morpho-syntactic features. Acquisition of a large bilingual parallel corpus for the desired domain and language pair requires a lot of time and effort. In this paper we report the performance comparison on training corpora of different sizes: 1k, 10k and 100k. Experiments were performed on small to middle-sized sentences of IJS-SVEZ corpus.

Strojno prevajanje iz slovenščine v angleščino s korpusi različnih velikosti in morfo-sintaktičnimi oznakami

Statistično strojno prevajanje na osnovi besed se kaže kot zelo obetavni pristop na področju strojnega prevajanja. Težavnost pregibnih jezikov je razpršenost podatkov. Delno jo rešujemo z večanjem korpusov za učenje in z uporabo dodatnih jezikovnih informacij: leme, in morfosintaktične oznake. V pričujočem članku analiziramo vplive različnih tipov jezikovnih informacij in različnih velikosti učnih korpusov. Pri eksperimentih smo uporabili IJS-SVEZ korpus.

1. Introduction

Research in statistical machine translation was pioneered at IBM (P. F. Brown and Mercer, 1993). They developed a language-independent framework, which was later re-implemented, improved, and the software has become freely available. Given these tools and a parallel corpus, a statistical machine translation system can be built in a relatively short time. The quality of the system closely depends on the features of the training corpus.

The historical enlargement of the EU has brought many new challenging language pairs for machine translation. A lot of work has been done on Czech (Čerjek et al., 2003), Polish (Jassem, 2004), Croatian (Brown, 1996), Serbian (Popović et al., 2004) and not at last Slovenian (Vičič and Erjavec, 2002; Romih and Holozan, 2002). This paper studies the translation direction Slovenian to English.

Acquisition of a large bilingual parallel corpus for the desired domain requires a lot of time and effort. Therefore, investigation of statistical machine translation with a small amount of training data is receiving more and more attention (Popović et al., 2004). In this paper we analyse statistical translation systems built on the largest Slovenian-English parallel corpus IJS-SVEZ (Erjavec, 2006). We analyse the results obtained with different amounts of training data, extracted from the same corpus.

2. Statistical Machine Translation

Statistical machine translation uses a notation of a source string $f_1^J = f_1 ... f_j ... f_J$, which is translated into a target string $e_1^I = e_1 ... e_i ... e_I$. In our experiments a source string is a Slovenian sentence and a target string is an English sentence. I is the length of the target string and J is the length of the source string. Among all possible target strings, the string with the highest probability as given by the Bayes' decision rule is chosen:

$$\hat{e}_1^I = \arg\max_{e_1^I} P(e_1^I|f_1^J) = \arg\max_{e_1^I} P(e_1^I) \cdot P(f_1^J|e_1^I) \quad (1)$$

$P(e_1^I)$ is the language model (of the target language) and $P(f_1^J|e_1^I)$ is the translation model. The $\arg\max$ operation denotes the search problem. In this paper we will focus on a translation model, which is based on an alignment model.

3. Translation Model

In the translation model the terms 'target language' and 'source language' are reversed. In the translation model the term 'target language' refers to the Slovenian language and the 'source language' refers to the English language. The translation model is based on word alignment. Given an English string e and a Slovenian string f, a word alignment is a many-to-one function that maps each word in f onto exactly one word in e, or onto the NULL word. The

NULL word is an invisible word in the initial position of an English sentence e_0. It accounts for Slovenian words that have no counterpart in the English sentence. More than one Slovenian word can be mapped onto the same English word. In the Slovenian string of words, we distinguish the heads from the non-heads. The head is the leftmost word of the group mapped to the same English word. All subsequent words in the same group are non-heads. A group of Slovenian words does not always contain neighbouring words. A sample of word alignment is shown in Figure 1. Each Slovenian word has its counterpart in an English sentence. Two Slovenian words ('Bil' and 'je') are mapped to the same English word ('was'). The word 'Bil' is a head word and 'je' is a non-head word. In this example, these two words are neighbouring words, but it is not always the case.

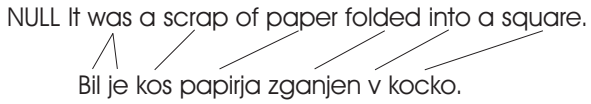

Figure 1: A sample alignment of sentence-pair.

An additional sample of word alignment is shown in Figure 2. The NULL word is an artificial construct in the initial position of an English sentence.

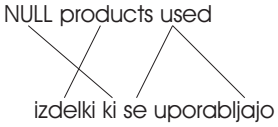

Figure 2: A sample alignment using a NULL word as a counterpart of Slovenian words that have no translation in English.

Word-for-word alignments of the translated sentences are not known. All possible alignments for a given sentence pair (e, f) are taken into account. An alignment for a sentence pair is denoted by a.

A series of five translation models (Model 1 to Model 5) were proposed by IBM (P. F. Brown and Mercer, 1993). Models 4 and 5 are the most sophisticated. We will focus on Model 4. Model 4 computes the probability $P(a, f|e)$ of a particular alignment and a particular sentence f given a sentence e. This probability is a product of five individual decisions:

$t(f_j|e_i)$ - translation probability. It is the probability of Slovenian word f_j being a translation of English word e_i.

$n(\phi_k|e_i)$ - fertility probability. An English word can be translated into zero, one or more than one Slovenian word. This phenomenon is modelled by fertility. The fertility $\phi(e_i)$ of an English word e_i is the number of Slovenian words mapped to it. The probabilities of different fertility values ϕ_k for a given English word are estimated.

p_0, p_1 - fertility probability for e_0. Instead of fertilities $\phi(e_0)$ of a NULL word, one single parameter $p_1 = 1 - p_0$ is used. It is the probability of putting a translation of a NULL word onto some position in a Slovenian sentence.

$d_1(\Delta j|\mathcal{A}(e_i), \mathcal{B}(f_j))$ - distortion probabilities for the head word. Δj is the distance between the head of current translation, and the previous translation. It may be either positive or negative. Distortion probabilities model different word order in the target language in comparison to the word order in the source language. Classes of words are used instead of words.

$d_{>1}(\Delta j|\mathcal{B}(f_j))$ - distortion probabilities for the non-head words. In this case Δj denotes the distance between the head and non-head word.

Model 4 has some deficiencies. Several words can lie on top of one another and words can be placed before the first position or beyond the last position in the Slovenian string. An empty word also causes problems. Training results in many words being aligned to the empty word. Model 5 is a reformulation of Model 4, in order to overcome some problems. An additional parameter is trained which denotes the number of vacant positions in the Slovenian string. It is added to the parameters of the distortion probabilities. In our experiments Models 4 and 5 will be trained, but only Model 4 will be used when decoding. Model 5 is not yet supported by the decoding program.

This was a short overview of the translation model. Readers interested in a more detailed description are referred to the paper (P. F. Brown and Mercer, 1993).

4. Adding Morphological Information

Previous work has shown that, for highly inflective languages, morphological information may be quite useful (Popović et al., 2004). The question arises, how much can be gained by adding morphological information.

A very basic way to modify input data using morphological information is by replacing each word-form with associated lemma. We expect this transformation would lead to an improvement in translation quality due to the restriction of data sparsity.

Since lemmatisation removes some useful information, we proceed by adding information from morpho-syntactic tags. These tags provide values along several morphological dimensions, such as part of speech, gender, number, etc. First only POS (Part Of Speech) tag is used, afterwards the complete MSD (Morpho-Syntactic Description) code is attached (Erjavec, 2004). In the latter case, data sparsity is increased because of homographs.

The translation model uses words grouped into classes. We analyse the influence of morpho-syntactic information on word grouping. The comparison was carried out between monolingual automatic clustering based on mutual information and clustering based on MSD codes.

In the following section four different sets of experiments are described, which differ in the ways the Slovenian lemma and morpho-syntactic tags are used.

The contribution of morphological information is closely related to the amount of training data and to its domain adequacy. We compare translation models, trained on different amounts of training data.

5. Experiments

5.1. SVEZ-IJS corpus

All experiments were performed on SVEZ-IJS corpus, a large parallel annotated English-Slovenian corpus. It contains approx. 10 million words of legal texts of the European Union, the ACQUIS Communautaire. The corpus is encoded in XML (according to TEI P4) and linguistically annotated at word-level. Tagging was performed by using TnT trigram tagger. Tagging accuracy for Slovenian was approx. 90%. CLOG (which is based on machine learning) was used for automatic lemmatisation. The estimated accuracy was approx. 95%. All corpus processing steps were performed by authors of the corpus and are described in some details in (Erjavec, 2006).

We discarded sentences longer than 15 words from the corpus, because of the computational complexity. The test set contained 25,000 sentences, taken at regular intervals from the corpus (homogeneous partition). The experiments were performed using three train sets, which differed in size (measured in sentences): 1k, 10k and 100k. There was no overlapping between the train and test sets. The vocabulary contained all units with occurance frequency (in the train set) greater than 2. All singletons (in training set) are mapped to the unique symbol UNK.

5.2. Tools

The experiments were performed using only publicly available third-party tools. The language model was trained by using the CMU-SLM toolkit (Rosenfeld, 1995). Classes of words were automatically created by means of the tool presented in (Maučec, 1997) and developed for language modelling. Translation model was trained using GIZA++ (Och and Ney, 2003). The decoding of test sentences was performed by the ISI ReWrite Decoder (Germann, 2003). Translations were evaluated using Word Error Rate (WER) and Bleu score (Papineni et al., 2001).

5.3. Translation model based on words

In our first set of experiments all word forms appeared as unique tokens and were exposed as candidates for word-to-word alignments. The Slovenian vocabulary (determined by the largest train set) contained 46,475 units (words). This vocabulary resulted in 5.0% OOV rate.

Before training, Slovenian words were mapped into 1000 classes and English words into 100 classes. A conventional trigram language model was built for the English language. The language model remained the same in all experiments. 10 iterations of training were performed for each translation model (1-5). The numbers of iterations were fixed for all experiments. Translation results are in Table 1. As expected, the error rate of the system trained on extremely small amounts of corpus is high. Using the 10-times larger train set the Bleu score improved by 89% relatively. When we used a train set of 100k sentences we obtained additional improvement of the Bleu score by 45%.

Train Set Size	WER [%]	Bleu [%]
1k	78.2	15.31
10k	61.0	28.92
100k	46.6	41.97

Table 1: Translation results. Translation model is based on word-forms.

5.4. Translation model based on lemmas

The purpose of the second set of experiments was the reduction of data sparsity. Here we used the lemmatised Slovenian part of the corpus. The English part remained unchanged. The Slovenian vocabulary (determined by the largest train set) contained 29,384 units (lemmas). The Slovenian vocabulary was reduced by 36% relatively (in comparison to the word-based translation model). This vocabulary resulted in a 2.7% OOV rate, which is 2.3% (absolute) lower than in the case of the word-based translation model. The translation results are in Table 2. A relative improvement is calculated to each value of evaluation metric (comparing the results with word-based baseline system). We achieved some improvements in the first two experiments, where data sparsity problem is more evident. In the last experiment we had worse results, because some information is lost by lemmatisation.

Train Set Size	WER [%]	Imp. [%]	Bleu [%]	Imp. [%]
1k	76.4	2.3	15.40	0.6
10k	59.3	2.8	30.41	5.2
100k	47.5	-1.9	41.36	-1.5

Table 2: Translation results. Translation model is based on lemmas.

5.5. Translation model based on lemmas and POS tags

We wanted to further examine the influence of morphosyntactic information is the translation process. Each Slovenian word was replaced by its lemma and the POS tag attached to it. The Slovenian vocabulary (determined by the largest train set) contained 30,450 units (lemmas with POS tag). This vocabulary resulted in a 2.9% OOV rate. Translation results are in Table 3. A relative improvement is calculated to each value of evaluation metric (comparing the results with word-based baseline system). In the first two experiments the improvement was not as evident as in the previous set of experiments with lemmas. In the last case (using 100k sentences in training) worsening of the Bleu score is smaller, because less information went astray.

Train Set Size	WER [%]	Imp. [%]	Bleu [%]	Imp. [%]
1k	76.3	2.4	15.38	0.5
10k	59.8	2.0	29.52	2.1
100k	47.7	-2.3	41.82	-0.4

Table 3: Translation results. Translation model is based on lemmas and POS tags.

5.6. Translation model based on lemmas and MSD codes

In this set of experiments we wanted to observe the influence of complete morpho-syntactic information. Slovenian words were replaced by lemmas and MSD codes were attached to them. The Slovenian vocabulary (determined by the largest train set) contained 59,339 units (lemmas, with MSD code). This vocabulary resulted in a 6% OOV rate.

In these experiments we expose the problem of homographs. For example the word *gori* can be replaced either by *goreti_*[VMIP3S–N] or by *gori_*[RGP], depending on the context. In addition, the problem of data sparseness increases. The translation results are in Table 4. The results

Train Set Size	WER [%]	Imp. [%]	Bleu [%]	Imp. [%]
1k	83.3	-6.5	10.35	-32.4
10k	66.5	-9.0	24.51	-15.2
100k	49.0	-5.1	40.60	-3.3

Table 4: Translation results. Translation model is based on lemmas and MSD codes.

were again compared against the word-based baseline system. We can see that using complete MSD code "adds a lot of noise" to the translation process. It should be noted that this observation depends tightly on the language pair under consideration and the direction of the translation. For example most MSD codes add usefull information to lemmas if we translate from one highly inflectional language to the other. The same is true, if we change the translation direction in our experiments.

5.7. Translation model based on word-forms and MSD classes

In the last set of experiments we used word forms as modelling units once again, but replaced automatic classes with classes based on MSD codes. Each distinct MSD code defines one class. All words having the same MSD code were mapped to the same class. The vocabulary size and OOV rate are the same as in first set of experiments (see Section 5.3.).

Train Set Size	WER [%]	Imp. [%]	Bleu [%]	Imp. [%]
1k	78.8	-0.8	15.42	0.6
10k	60.9	0.2	30.56	5.7
100k	47.1	-1.2	42.55	1.3

Table 5: Translation results. Translation model is based on word-forms and MSD classes.

The translation results are in Table 5. Comparing Bleu scores against the word-based baseline system shows, that MSD codes contain some information about word reordering between the source and target languages.

6. Conclusion

This paper reports our first experiments using SVEZ-IJS corpus. We were interested in the influence of morpho-syntactic information on statistical machine translation using different amounts of training data. Lemmatisation reduces data sparsity significantly and improves the results when using small training corpus. In the case of a large training corpus the performance deteriorated, because some useful information was lost. Using complete morpho-syntactic information is unwise choice due to the increase in data sparsity. It seems that only a subset of morpho-syntactic features is important, which depends on the language pair under consideration. Our future work will proceed in the direction of extracting useful morpho-syntactic features by a data driven approach.

7. References

R. Brown. 1996. Example-based machine translation in the Pangloss system. *In Proceedings of COLING-96*.

M. Čerjek, J. Cuřin, and J. Havelka. 2003. Czech-English dependency-based machine translation. *In Proceedings of the European Chapter of the ACL*, Vol. 1.

T. Erjavec. 2006. The English-Slovene ACQUIS corpus. *In Proceedings of the conference LREC*, pp.: 2138–2141.

T. Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *In Proceedings of the conference LREC*, pp.: 1535–1538.

U. Germann. 2003. Greedy Decoding for Statistical Machine Translation in Almost Linear Time. *In Proceedings of the HLT-NAACL-2003*. URL: http://www.isi.edu/licensed-sw/rewrite-decoder/.

K. Jassem. 2004. Applying Oxford-PWN English-Polish dictionary to machine translation. *In Proceedings of the 9th EAMT Workshop*.

M. S. Maučec. 1997. Statistical language modeling based on automatic classification of words. *In Proceedings of the workshop: Advances in speech technology*.

F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*. URL: http://www.fjoch.com/GIZA++.html.

V. J. D. Pietra P. F. Brown, S. A. D. Pietra and R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2001. Bleu: a Method for Automatic Evaluation of Machine Translation. RC 22176(W0109-022), IBM Research.

M. Popović, S. Jovičić, and Z. Šarić. 2004. Statistical Machine Translation of Serbian-English. *In Proceedings of the SPECOM-2004*.

M. Romih and P. Holozan. 2002. Slovensko-angleški prevajalni sistem. *In Proceedings of the conference Jezikovne tehnologije*.

R. Rosenfeld. 1995. The CMU Statistical Language Modeling Toolkit, and its use in the 1994 ARPA CSR Evaluation. *In Proceedings of the ARPA SLT Workshop*. URL: http://www.speech.cs.cmu.edu/SLM/toolkit.html.

J. Vičič and T. Erjavec. 2002. Vsak začetek je težak : avtomatsko učenje prevajanja slovenščine v angleščino. *In Proceedings of the conference Jezikovne tehnologije*.