Statistical machine translation using the IJS-ELAN Corpus

Mirjam Sepesy Maučec and Zdravko Kačič

Faculty of electrical Engineering and Computer Science

Smetanova 17, 2000 Maribor, Slovenia

E-mail: mirjam.sepesy@uni-mb.si, kacic@uni-mb.si

Abstract

In this paper, we describe our experiments on statistical machine translation from Slovenian to English performed on IJS-ELAN bilingual corpus. Four different experiments are reported. They differ from each other in the type of information used in training the translation model. In first experiment word forms of sentence-aligned corpus are used as modelling units. Second experiment uses lemmas instead of word forms in Slovenian part of the corpus. In the third experiment word forms are replaced by lemmas and MSD codes attached to them. Fourth experiment tries to combine the advantages of previously performed ones. Only publicly available tools are used for training language model and translation model, as well as for decoding the test set. The results are evaluated by automatically calculated WER.

1. INTRODUCTION

The paper discusses corpus-based machine translation. The use of corpora of bilingual parallel texts seems to offer a promising tool for the future, thanks to the progress that has been made in the field of statistical machine translation. Various methods have been proposed for processing the different levels of correspondence between two texts, an original and its translation. In this paper well grounded methods of statistical machine translation are tested on the Slovenian-English language pair.

To our knowledge, "pure" statistical machine translation of Slovenian language has not been widely studied yet. There exist only one thesis (Vičič, 2002), which gave us fist impression about the topic

2. STATISTICAL MACHINE TRANSLATION

The problem is to find the optimal English translation $\hat{e}$ of a Slovenian sentence f. Statistical machine translation is referred to as the noisy channel model:

$$\hat{e} = \arg\max_{e} P(f|e) \cdot P(e) \qquad (0.1)$$

$P(f|e)$ denotes the translation model and $P(e)$ the language model of English language. The search for the optimal translation is encompassed by $\arg\max_{e}$. Language model is a conventional trigram model with Katz back-off smoothing. This paper deals with translation model.

3. TRANSLATION MODELS

Translation model defines the correspondence between the words of the target sentence and the words of the source sentence. Translation model is a generative model, because it is a theory how Slovenian sentences are generated. First an English sentence is generated, and then it gets turned into a Slovenian one. Although we are building a Slovenian-to-English machine translation system, we reason in the opposite direction when training the translation model.

Translation models, defined by IBM in the early nineties, are used in our research. They are well documented (Brown at all, 1993) and the software for training them is publicly available (Och & Ney, 2003). The models are indexed from 1 to 4 according to their increasing complexity in training. The parameters are transferred from one model to another, for example from Model 2 to Model 3. It means that final parameter values of Model 2 are the initial parameter values of Model 3. Models 4 and 5 are the most sophisticated.

3. MODEL 4

Translation model is a word-for-word alignment model between two strings, an English string $e_1^I = e_1, ..., e_I$ and a Slovenian string $f_1^J = f_1, ..., f_J$. Because alignments are not known, the probability $P(a, f|e)$ for each particular alignment a is computed.

Model 4 is the final model, which is used by the decoder. It is composed of the following probabilities:

- $P(f_j|e_i)$ - translation probability. It is the probability of Slovenian word f_j being a translation of English word e_i.
- $P(\phi_k|e_i)$ - fertility probability. An English word can be translated into zero, one or more than one Slovenian word. This phenomenon is modeled by fertility. The fertility $\phi(e_i) = \phi_k$ 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 trained.

- p_0, p_1 - fertility probability for e_0. The word e_0 is an invisible word in the initial position of an English sentence. It accounts for Slovenian words that have no counterpart in the English sentence. Instead of fertilities $\phi(e_0)$ one single parameter $p_1 = 1 - p_0$ is used. It is the probability of putting a translation of word e_0 onto some position in a Slovenian sentence.

  Between words being a translation of the same word we distinguish a head word and non-head words. Head word is the first word in the translation. All other words are non-head words.
- $P_{=1}\left(\Delta j \middle| \mathcal{A}(e_i), \mathcal{B}(f_j)\right)$ - 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. $\mathcal{B}$ denotes mapping into classes for Slovenian words and $\mathcal{A}$ for English words.
- $P_{>1}\left(\Delta j \middle| \mathcal{B}(f_j)\right)$ - distortion probabilities for 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 the problems. Training results in many words aligned to the empty word. Model 5 is a reformulation of Model 4 in order to overcome some problems. An additional parameter is trained. It 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.

4. IJS-ELAN CORPUS

The translation system was tested on the IJS-ELAN corpus (Erjavec, 2002)[1]. The corpus has parts, which have a Slovenian origin and an English translation, and parts with origins in English and translations in Slovenian. The corpus is encoded in XLM/TEI P4. It is aligned at the sentence level, tokenised, and the words are annotated with disambiguated lemmas and morpho-syntactic descriptions (MSD). All annotations were done by the authors of the corpus.

We observed that the English part contains 18% more words than the Slovenian part. The average English sentence is 3 words longer than the average Slovenian sentence. One reason lies in determiners and pronouns. The subject pronouns in English (I, he, they) usually have a zero form in Slovenian.

The Slovenian corpus contains twice as many unique words than the English corpus; this is because of the highly inflectional nature of the Slovenian language. Almost half the words are singletons (they appeared only once in the training corpus). The data exposes the problem of data sparsity of the corpus and indicates the difficulty of the translation process.

5. EXPERIMENTS

5.1 TRAIN AND TEST SET

We discarded sentences longer than 15 words from the corpus because of the computational complexity. The rest of the corpus was split into training and test sets in the ratio 8:2. The test sentences were taken at regular intervals from the corpus (homogeneous partition). Some statistics of the training corpus are collected in Table 1. The training set contained 12,044 sentence pairs. Each appearance of a word or any other string of characters between two spaces is counted as one unit. The Slovenian part was 86,036 units long and the English part contained 97,062 units. The test set consisted of 3,069 sentences.

The vocabulary contained all those words (units), which appeared in the training set or in the test set. Almost half of the vocabulary units were singletons. Zerotons are units, which do not appear in the training corpus, but occur in the test set. These units not only remained untranslated but also "added noise" to the translation process of other words.

5.2 TOOLS

All experiments have been performed using only publicly available third-party tools. The language model was made by using the CMU-SLM toolkit (Rosenfeld, 1995). Classes of words were made by the tool, developed for language modelling (Maučec, 1997). The translation model training was performed using a program GIZA++ (Och, 2003). The decoding of test sentences was performed by an ISI ReWrite Decoder (Germann, 2003). Translations were evaluated using Word Error Rate metric (WER).

5.3 FIRST EXPERIMENT

In our first experiment all word forms appeared as unique tokens and were exposed as candidates for word-for-word alignments.

Before training, Slovenian words were mapped into 1000 classes and English words into 100 classes. The numbers of classes were predefined and were chosen to be the same as the number of different MSD codes in the corpus.

For English language a conventional trigram language model was built with Good-Turing discounting for bigrams and trigrams with counts lower than 7. No n-grams were discarded. Training corpus was the whole English part of IJS-ELAN corpus. It is relatively small, so there are a lot of singletons with significant information. The language model perplexity of the test set was 48.

[1] The authors are thankful for giving the corpus freely available.

	SLO part	ENG part
Sentences	12,044	
Units	86,036	97,062
Vocabulary	22,055	12,715
- singletons	11,355	5,611
- zerotons	2,566	1,274

Table 1: Training corpus for the first experiment.

For each translation model (Model 1 – Model 4) 10 iterations were performed.

After training some interesting observations were made (see Figure 1). Although the training-set perplexity continuously decreased, the test-set perplexity jumped at each transition point. At the transition point, the final estimates of one model initialized the estimates of the next model. In subsequent iterations after transition points, the test-set perplexity slowly increased, especially in Models 1 and 4. Each iteration of Model 4 made the test-set perplexity worse. The only exception was the transition to Model 5, although the estimates never improved the estimates obtained at the beginning of the training. The same observations have also been reported in Czech-English experiments (Al-Onaizan at all, 1999). We speculated that the reason was the small size (and consequently data sparsity) of the training corpus, so the translation probabilities become over-trained. Better alignments for training-set did not lead to better translations of previously unseen test-set.

In the first experiment only 37.5% of words got the correct translation. We obtained WER=71,6% (see first row in Table 4).

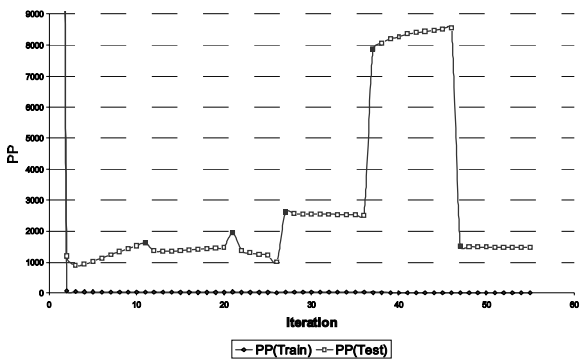

Figure 1: Train set and test set perplexities in first experiment

5.4 SECOND EXPERIMENT

The purpose of the second experiment was the reduction of data sparsity.
In the second experiment we used lemmatised Slovenian part of the corpus. English part remained unchanged. Lemmatising the Slovenian corpus reduced the data sparsity to a great extent (see Table 2).

	SLO part	ENG part
Sentences	12,044	
Units	86,036	97,062
Vocabulary	12,629	12,715
- singletons	5,654	5,611
- zerotons	1,292	1,274

Table 2: Training corpus for the second experiment.

New clustering of Slovenian lemmas was performed. The Slovenian lemmas were automatically clustered into 100 classes. Because Slovenian part of the corpus was lemmatised, there was no need to use extended set of classes.

The GIZA++ training was repeated. Comparison of train-set and test-set perplexities confirmed our assumptions about Model 4 over-training in first experiment (see Figure 2). In the second experiment each iteration of Model 4 training made a slight reduction of the test-set perplexity. Transition to Model 5 brought further improvements.

In the second experiment we obtained WER=68.5% (see second row in Table 4). Although some information was lost by lemmatization, the data sparsity reduction improved the results.

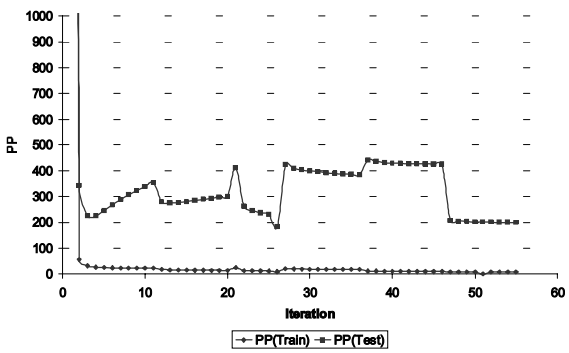

Figure 2: Train set and test set perplexities in second experiment

5.5 THIRD EXPERIMENT

In the third experiment we want to examine what is the influence of morpho-syntactic information on the translation success.

Slovenian words were replaced by lemmas and MSD codes attached to them. In this experiment we expose the problem of homographs. For example the word gori was replaced by goreti_[VMIP3S–N] and by gori_[RGP]. In this experiment we increase the data sparsity, so worsening of the translation results was expected. The data sparsity of third experiment is evident from Table 3.

	SLO part	ENG part
Sentences	12,044	
Units	86,036	97,062
Vocabulary	27,030	12,725
- singletons	14,809	5,645
- zerotons	3,409	1,272

Table 3: Training corpus for the third experiment.

Two versions of third experiment were performed. In the first one classes were made automatically. In this version of the experiment we obtained WER=75.5% (see third row in Table 4).

In the second version of third experiment classes of words were defined by MSD codes. We had 1100 different MSD codes. In this case we obtained WER=71.5% (see fourth row in Table 4).

The results of third experiment showed the problem of data sparsity. Data sparsity is even greater than shown at the beginning. The results obtained with MSD codes gave us the anticipation that MSD code is well correlated with the position of word in sentence.

5.6 FOURTH EXPERIMENT

In the fourth experiment the influence of MSD codes was further examined. We used them to improve distortion probabilities of our first experiment. The experimental setup remained unchanged, only automatic classes were replaced by MSD classes. Distortion probabilities of words depended directly on morpho-syntactic features of words. In this experiment we obtained WER=69.9% (see fifth row in Table 4).

One additional version of this experiment was performed. In this experiment translation probabilities $P(f_j|e_i)$ were imported from the second experiment, where they were learned on lemmas. Each lemma-based probability value was assigned to all word forms, which belong to that lemma. The probabilities were normalized afterwards.

In the last experiment we obtained WER=68.9% (see last row in Table 4). This result is only slightly worse than the result obtained by lemmas. This experimental setup did not need the lemmatiser in decoding phase.

	CORR (%)	WER (%)
1. EXP.	37.5	71.6
2. EXP.	40.6	68.5
3. EXP. (a)	33.0	75.5
3. EXP. (b)	35.9	71.5
4.EXP. (a)	38.9	69.9
4.EXP. (b)	40.0	68.9

Table 4: Final results of all experiments

6. CONCLUSION

In this paper, we have discussed different types of translation model. The problem of data sparsity was outlined.

From the experiments we concluded that using lemmas is a good starting point for data sparsity reduction. On the other hand it has been shown that MSD codes are of great value. In the future we will examine how to use them more reasonable. Not all the information in MSD codes is important for translation. We would like to reduce its content just to the important parts.

7. References

Al-Onaizan, Y., Curin, J., Jahr M., Knight K., Lafferty, J., Melamed D., Och F. J., Purdy D., Smith N. A., Yarowsky D. (1999). Statistical Machine Translation. Final report, JHU Workshop.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J. And Mercer, R. L. (1993). The mathematic of statistical machine translation : Parameter estimation.. Computational Linguistic, 19(2):263:311.

Erjavec, T. (2002). Compiling and Using the IJS-ELAN Parallel Corpus. Informatica, Vol. 26.

Germann, U. (2003). Greedy Decoding for Statistical Machine Translation in Almost Linear Time. In Proceedings of HLT-NAACL-2003, Edmonton, AB, Canada.

Maučec, M. S. (1997). Statistical language modeling based on automatic classification of words, In Proceedings of workshop : Advances in Speech Technology, Maribor : Faculty of Electrical Engineering and Computer Science.

Och, F. J. & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1).

Rosenfeld, R. (1995). The CMU Statistical Language Modeling Toolkit, and its use in the 1994 ARPA CSR Evaluation. In Proceedings of the ARPA SLT Workshop, Austin, TX.

Vičič, J. (2002). Statistično strojno prevajanje naravnih jezikov, Master thesis.