View metadata, citation and similar papers at core.ac.uk brought to you by CORE

Journal of Computing and Information Technology - CIT 15, 2007, 1, 47–59
doi:10.2498/cit.1000760

Statistical Machine Translation from Slovenian to English

Mirjam Sepesy Maucec and Zdravko Kacic

Faculty of Electrical Engineering and Computer Science, University of Maribor

In this paper, we analyse three statistical models for the machine translation of Slovenian into English. All of them are based on the IBM Model 4, but differ in the type of linguistic knowledge they use. Model 4a uses only basic linguistic units of the text, i.e., words and sentences. In Model 4b, lemmatisation is used as a preprocessing step of the translation task. Lemmatisation also makes it possible to add a Slovenian-English dictionary as an additional knowledge source. Model 4c takes advantage of the morpho-syntactic descriptions (MSD) of words. In Model 4c, MSD codes replace the automatic word classes used in Models 4a and 4b. The models are experimentally evaluated using the IJS-ELAN parallel corpus.

Keywords: statistical machine translation, translation model, lemmatisation, morpho-syntactic description, Slovenian language.

1. Introduction

Machine translation (MT) from one human language to another is a longstanding goal of computer science. Statistical data analysis has been a minority approach in this field for a long time. The growing availability of bilingual, machine-readable text has stimulated interest in statistical methods. The use of statistical methods was first published by IBM in the early nineties [2]. The statistical machine translation (SMT) system constructs a general model of the translation process. It acquires specific rules automatically from bilingual and monolingual text corpora. A number of SMT systems have been presented over recent years. They share the same basic underlying principles (will be defined in Section 1.2), but differ in the structures and sources of their translation models. Some of them use word-based translation models [18, 16, 4, 15]. More sophisticates use complex phrase structures [27].

The historical enlargement of the EU has brought many new and challenging language pairs for machine translation. A lot of work has been done on Czech [4], Polish [12], Croatian [3], Serbian [20] and not at last Slovenian [6].

The Czech-English machine translation system is based on dependency trees. Dependency trees represent the sentence structure, as concentrated around the verb. The presented system was outperformed by the statistical translation system GIZA++/ISI ReWrite Decoder [16, 10, 18], trained on the same corpus.

The Polish-English MT system[12] uses an electronic dictionary annotated for morphological, syntactic, and partly semantic information. The dictionary is based on the corpus of Polish texts from the domain of computer science.

The Croatian language was used as one of the target languages in multilingual example-based MT [3]. Examples to be used were selected based on string matching, and inflectional and other heuristics, with no deep structural analysis.

Statistical machine translation of Serbian to English was also studied [20]. It was reported that reducing Serbian words into stems decreases error rates for the translation.

A Slovenian-English translation system called Presis has been developed by Amebis, d.o.o. company [21]. It is a rule-based MT, which includes translation memory technology. Rule-based MT analyses the morphology, syntax and semantic of a source sentence according to previously defined rules. Translation memory enhances the human translation effort. It stores the source and target language strings of words translated by the translator in a database. These strings are much shorter than sentences, usually two or three words in length. A tool was developed, being a translation aid for human translators [22]. The translation output given by a system is reviewed and completed by the human translator. As the translation effort progresses, the translation memory grows.

To our knowledge, pure statistical machine translation from/in the Slovenian language has only been published once within a very limited scope[26]. SMT cannot compete against rule-based systems with translation memory. SMT is language-independent and uses bilingual corpus as the only knowledge source. Consequently, the results are much worse, but they add some valuable observations about the translation process. It has been shown that SMT and rule-based MT can be successfully combined in a hybrid approach [5].

In this paper we present our first experimental results in SMT. They are intended to be used in our speech-to-speech translation (SST) system, which is still under development. It consists of three complex components: speech recognition (converts speech to text), translation (translates text in one language to text in another language) and speech synthesis (converts text to speech). To date only the language resources used in our SST system have been published [25].

Our idea in this paper is to use bilingual corpus as the sole knowledge source of an MT system and to avoid language specific rules. Statistical methods are the path to follow.

1.1. Overview

In Section 1.2, we present a short review of the statistical machine translation models. In Section 2 we describe the alignment model, named IBM Model 4. Section 3 introduces lemmatisation, which yields significantly better results, especially for a small training corpus. The use of morpho-syntactic classes is briefly described in Section 4. In Section 5 we give some data about IJS-ELAN Corpus. Section 6 describes two well known evaluation criteria, which have been used in our experiments. The experiments are discussed in Section 7.

1.2. Statistical Machine Translation

The following mathematical notation of the statistical machine translation, taken from the paper [2], will be used: a source string of words $f = f_1 f_2 ... f_j ... f_J$ is to be translated into a target string of words $e = e_1 e_2 ... e_i ... e_I$. The string with the highest probability is chosen from among all possible target strings, as given by the Bayes decision rule:

$$e = \arg\max_e P(e|f) = \arg\max_e P(e)P(f|e) \tag{1}$$

$P(e)$ is the language model of the target language, whereas $P(f|e)$ is the translation model. The language and translation models are independent knowledge sources. The arg max operation denotes the search for an output string in the target language.

All SMT systems share these underlying principles for applying a translation model in order to capture the lexical and word reordering relationships between two languages. They are complemented by a target language model to drive the search process through translation model hypotheses.

Language model is essentially the same as that for speech recognition and has been dealt with elsewhere in that context [14, 23]. Search process is also not our topic. It is described in [10].

This paper concentrates on the translation model. A translation model is a generative model because it is a theory as to how Slovenian sentences are generated. Although we are building a Slovenian-to-English machine translation system, we reason in the opposite direction when training the translation model (see $P(f|e)$ in Equation 1).

A series of five translation models (Model 1 to Model 5) were proposed by IBM in the nineties [2] and remain the most attractive to this day [16]. They are based on word alignments and are indexed according to their increasing training complexity. The parameters are transferred from one model to another, for example from Model 2 to Model 3. It means that the final parameter values of Model 2 are the initial parameter values of Model 3. Models 4 and 5 are the most sophisticated. We will focus on Model 4 and, partly, on Model 5.

One major disadvantage of single-word-based approaches is that contextual information is not taken into account. The translation model probabilities are based only on single words. For many words, the translation depends heavily on the surrounding words. This problem is partly addressed by the language model.

2. IBM Model 4 (4a)

In this section the translation model is described. We focus on Model 4 although some observation will also be made of Model 5.

In our 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. More than one Slovenian word can be mapped onto the same English word (see Figure 1).

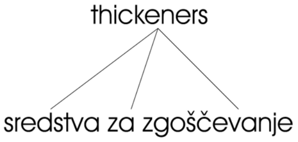

Fig. 1. Many-to-one mapping (example taken from anx2.en.77 and anx2.sl.77).

Each word in f is mapped onto exactly one word in e, or onto the NULL word e_0. The NULL word is an artificial construct in the initial position of an English sentence. It accounts for Slovenian words that have no counterpart in the English sentence (see Figure 2).

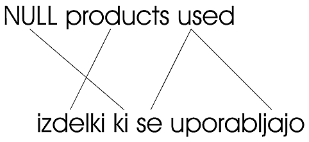

Fig. 2. NULL word as counterpart of Slovenian words that have no translation in English (example taken from anx2.en.203 and anx2.sl.203).

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. In Figure 1 the Slovenian word 'sredstva' is a head word and words 'za' and 'zgoščevanje' are non-head words. A group of Slovenian words does not always contain neighbouring words.

An additional sample of word alignment is shown in Figure 3. 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, the head word and non-head word are neighboring words, but this is not always the case.

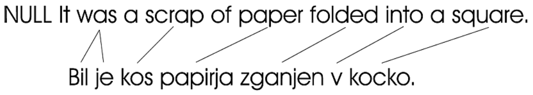
NULL It was a scrap of paper folded into a square.
Bil je kos papirja zganjen v kocko.

Fig. 3. A sample alignment of sentences orwl.en.2018 and orwl.sl.2018 from the IJS-ELAN corpus.

Word-for-word alignments of the translated sentences are unknown. All possible alignments for a given sentence pair (e,f) are taken into account. An alignment for a sentence pair is denoted by a.

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 trained.

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 the 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 words.

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 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.

This was a short overview of the translation model. Readers interested in a more detailed description are referred to the paper [2].

2.1. Automatic Word Classes

Word mapping into classes is performed before training the distortion probabilities. The grouping of words into classes is based on the assumption, that word displacement depends on certain features, which are common to many words. We construct two independent mappings, $\mathcal{A}$ for English words and $\mathcal{B}$ for Slovenian words. f_1^M denotes the Slovenian part of the bilingual corpus and M is its length. We have the following probability model:

$$p(f_1^M|\mathcal{B}) = \prod_{j=1}^{M} p(\mathcal{B}(f_j)|\mathcal{B}(f_{j-1})) \cdot p(f_j|\mathcal{B}(f_j)) \tag{2}$$

A maximum likelihood approach is performed in order to determine the optimal classes $\hat{\mathcal{B}}$ for a given number of classes:

$$\hat{\mathcal{B}} = \arg\max_{\mathcal{B}} p(f_1^M|\mathcal{B})$$

Words are mapped into classes so as to preserve mutual information between adjacent classes in a corpus. It is necessary to determine the number of classes in advance, as the optimum is reached if every word is in a class of its own. Such word classes are useless. The aim of word classes is to solve the problem of sparse data. The number of distortion probabilities is reduced. We use the implementation of an optimization algorithm used for language modelling [14], where we face the same problem of data sparsity. Having classes, bigrams of words are replaced by bigrams of classes. For example, if we use the mapping of 20 000 words into 100 classes, the max. number of word-bigrams is $20\,000^2$ and the max. number of class-bigrams is 100^2.

Having described the formation of word classes, we have embraced all ideas about IBM Model 4, (our Model 4a).

3. Lemmatisation (Model 4b)

The Slovenian language as a highly inflectional language needs some additional linguistic preprocessing. Lemmatisation is introduced as the first type of transformation, used in the Slovenian part of the parallel corpus.

Lemmatisation represents a normalisation step, where all inflected forms of a word are reduced to its common lemma. The term 'lemma' means a word form in its canonical form (e.g., infinitive for verbs, nominative singular for regular nouns, etc.). We distinguish between a lemma and the stem of a word form. For example, the feminine noun dual *postelji*, has *postelja* (eng. bed) as its lemma, *postelj-* as its stem and $-i$ as the inflectional ending. Lemmatisation consists of morphological analysis, to identify the ending and isolate the stem and synthesis, to join the canonical ending to it. Lemmatisation based on a single word is ambiguous. For example, the word *postelji* can have two different lemmas, *postelja* and *postlati* (eng. to make the bed). Unambiguous lemmatisation of words is only possible if the words in corpus are morphologically analyzed. Lemma is produced from a word form given its morpho-syntactic description (MSD). For example, the word *postelji* has *postelja* as its lemma, if it has MSD `Ncfsd`, `Ncfsl`, `Ncfdn` or `Ncfda`, and `postlati`, if it has MSD `Vmp3s-a------e`. Morpho-syntactic description will be discussed in Section 4.

Lemmatisation was not part of our work. It has been done by the authors of the corpus [8]. Only the lemmatised Slovenian part of the corpus is used in our experiments. The English counterpart is used in basic unlemmatised form. When using lemmatisation, the frequency statistics of Slovenian and English parts become similar. This observation will be presented in Section 7. On the other hand, using lemmatisation we discard information about number, tense, gender and other features which are, in some cases, relevant for English translation.

The model with a lemmatised Slovenian corpus is referred to as Model 4b. In Model 4b, f denotes a Slovenian lemma. In this model, the automatic clustering of Slovenian lemmas has to be performed.

3.1. The Slovenian-English Dictionary

Intuitively, the bilingual dictionary could improve the translation model, essentially the probability $t(f_j|e_i)$. The use of a dictionary changes the co-occurrence counting in the first iteration of Model 1. In parallel sentences $f = f_1f_2...f_j...f_J$ and $e = e_1e_2...e_i...e_I$, a pair (f_j, e_i) is counted as a co-occurrence pair if one of the following two conditions is met:

- f_j and e_i occurs as an entry in the dictionary, or
- f_j does not occur in the dictionary with any e_i $(i \in \{1, I\})$ and e_i does not occur in the dictionary with any f_j $(j \in \{1, J\})$.

The use of a dictionary improves, not only the alignment of the words in the dictionary but also indirectly for other words.

Dictionary entries, which are not seen in the training corpus, fall out of use during the training, according to the previous two conditions. Some entries are still of great value. They contain alignments for vocabulary words which appear in the translation process but have not been seen during the training. Without the proper use of a dictionary, these words remain untranslated and take the wrong positions in the target sentence. In order to obtain full value from the dictionary, all entries covered by the vocabularies should be added to the training corpus.

In our work, a dictionary was used in one version of Model 4b. The experiments with Model 4b were split in two experiments: one only with lemmatisation (Model 4b-1) and the other with the dictionary as well (Model 4b-2). Having the Slovenian part of the corpus lemmatised, looking-up in the dictionary was straightforward. The dictionary was extended to all word forms associated with the English lemma, because the English part of the corpus was used in non-lemmatised form. Only parts of the dictionary were used. We limited the use of entries to those where both the Slovenian lemma and the English word appeared in the corpus.

4. Morpho-syntactic Classes (Model 4c)

In Section 3 we have introduced the morpho-syntactic description (MSD), as it is used as a basis for lemmatisation. An MSD code is attached to each word in a corpus. It was defined in the MULTEXT-East project [7]. The MSD encodes the part-of-speech of the word in question and the values for additional attributes defined for each part-of-speech. For example, in the Slovenian part of the corpus the most common MSD is `Ncfsn`. It expands to part-of-speech:noun (`N`), type:common (`c`), gender:feminine (`f`), number:singular (`s`) and case:nominative (`n`). A big difference between languages is evident from the number of different MSDs. The English part of the corpus contains 134 MSDs and the Slovenian part 1100 MSDs.

The displacement of a word from its position in the source sentence, to its position in the target sentence depends on its syntactic features. The displacement information could be best inferred from a comparison of syntactic parse trees [4]. The only information we have relates to morpho-syntactic descriptions. So we group words into classes based on their MSD codes. The automatic word classes (used in Model 4a and Model 4b) are replaced by MSD classes. The distortion probabilities are then obtained, based on MSD classes. The resulting translation model is called Model 4c.

5. IJS-ELAN Corpus

The translation system was tested on the IJS-ELAN corpus [6]. The corpus has parts which have a Slovenian origin and an English translation, and parts with origins in English and translation in Slovenian. In spite of linguistic differences, we use all the parts in the same way. Half of the corpus contains documents produced by the Slovenian government. The remaining part: two texts, which deal with computers, one is about pharmaceuticals, and one is a literary work. All these collections are examples of written language, except one, which contains speeches by the former President of Slovenia. The corpus is encoded in XML/TEI P4. It is aligned at the sentence level, tokenised, and the words are annotated with disambiguated lemmas and morpho-syntactic descriptions (MSD) [6]. It is reported that Amebis d.o.o. company performed lexical annotation and that the TnT tagger was used afterwards to solve the ambiguity. The authors of the project reported 93% tagging accuracy.

An example sentence from the corpus is given in Figure 4.

Some corpus statistics are collected in Table 1. The exact values are given throughout the paper to encourage any interested reader to repeat (and/or improve) the experiments. The resources used are widely available and we only use publicly available third-party tools.

```
<seg id="ekol.sl.1676" corresp="ekol.en.1676">
<w ana="Ncmpg" lemma="podatek">Podatkov</w>
<w ana="Spsa" lemma="za">za</w>
<w ana="Npfsa" lemma="slovenija">Slovenijo</w>
<w ana="Q" lemma="&scaron;e">&scaron;e</w>
<w ana="Vmip1p--y" lemma="imeti">nimamo</w>
<c ctag=".">.</c>
</seg>

<seg id="ekol.en.1676" corresp="ekol.sl.1676">
<w ana="Ncnp" ctag="NNS NNS" lemma="figure">
Figures</w>
<w ana="Sp" ctag="IN IN" lemma="for">for</w>
<w ana="Np" ctag="?NN NP" lemma="slovenia">
Slovenia</w>
<w ana="Vmip-p" ctag="VBP BER" lemma="be">
are</w>
<w ana="Rmp" ctag="RB XNOT" lemma="not">not</w>
<w ana="Rmp" ctag="RB RB" lemma="yet">yet</w>
<w ana="Afp" ctag="JJ JJ" lemma="available">
available</w>
<c ctag=".">.</c>
</seg>
```

Fig. 4. Sentence pair from the IJS-ELAN corpus.

	Slo	Eng
Sentences	31 900	
Aver. sentence length	15.72	18.51
Tokens	498 906	587 481
– types	50 331	24 382
– singletons	24 830	10 575

Table 1: IJS-ELAN corpus.

It is interesting to note 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. This is called pronoun-dropping. The Slovenian corpus contains twice as many unique words than the English corpus, this is because of the highly inflective nature of the Slovenian language. Almost half of the words are singletons (they appeared only once in the training corpus). This fact indicates the difficulty of the translation process.

5.1. Training and Testing Sets

We discarded sentences longer than 15 words because of 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. Corpus division was obtained by using the Whittle program [11] with the following parameters: baseline ratio: 0.2, interleave ratio: 1 and sentence length restriction: 15.

The training set contained 12 064 sentences. The Slovenian part was 86 177 words long and the English part contained 97 258 words. The test set consisted of 3 123 sentences. Some statistics from the training corpus are collected in Table 2.

The vocabulary contained all the words 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

	Slo Words	Eng Words	Slo Lemmas
Sentences	12 064		
– units	86 177	97 258	86 177
Vocabulary size	22 072	12 725	12 643
– singletons	11 411	5 645	5 636
– zerotons	2 564	1 272	1 289
Co-occurence pairs	857 770		
– unique	486 312		413 696
– singletons	385 451		312 333

Table 2: Training corpus.

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. The statistics for the lemmatised Slovenian corpus are given in the third column. The average counts for words and lemmas were compared. Average count for Slovenian word was 9.97 and of lemma 19.94.

6. Evaluation Criteria

We used automatically computable metrics for the evaluation of translations. The first evaluation was performed after training, and before decoding of the test set. We measured the train-set perplexity $(PP(Train))$ and test-set perplexity $(PP(Test))$ of the translation model. Perplexity measures how well a translation model fits the (training/test) data. It is a function of probability. The translation model assigns a probability $P(f\,|e)$ to any pair of sentences f and e. Train-set perplexity $PP(Train)$ is computed as

$$PP(Train) = 2^{-\frac{\sum_{(e,f)\in Train} \log P(f\,|e)}{N}}$$

$Train$ denotes sentence-pairs from the training set. Test-set perplexity $PP(Test)$ has an analog definition.

In all translation experiments, we used the following two error criteria:

- WER (word error rate)[9]. It is computed as the minimum number of substitutions, insertions and deletions that have to be performed to convert the hypothetical into the reference sentence. WER is expressed in percentage.
- BLEU (bilingual evaluation understudy) [19, 1]. Word order in Slovenian sentences is quite relaxed.

As a result the word order in a hypothetical sentence can be different from that of a reference sentence, but nevertheless acceptable. We compute Bleu metric to partly overcome the problem of WER measurement. Bleu compares n-grams of both sentences and counts the number of matches. Its value ranges from 0 to 1. Only a translation identical to a reference translation will attain a score 1. It should be noted that human translators are usually scored as approx. 0.35 [19].

7. Experimental Results

The experiments were carried out for the translation direction Slovenian to English. The translation model training was performed using the program GIZA++ [16, 17]. IBM Models 1-4 were used as stepping stones. For example, the final estimates of Model 1 were used as initial estimates of Model 2. HMM model should be seen as a link between Model 2 and Model 3. 10 iterations for each Model were performed in all experiments. Model 5 was also trained, although it was not used in decoding. It was kept in training schema, because it improved Model 4 probabilities. Model 5 training yielded some interesting observations, which will be discussed in Section 7.3. The decoding of test sentences was performed using the ISI ReWrite Decoder [10, 13], which supports only Model 4.

Experiments were performed on Pentium IV 2.4 GHz with 2GB RAM, which runs SUSE Linux 8.2. Average training took 4 hours. Most of the time was spent on building automatic classes. Average decoding of the whole test set lasted for 5.6 mins.

7.1. Language model

All models use the same language model for the English language. The whole English part of the IJS-ELAN corpus was used for training. The language model was made by using the CMU-SLM toolkit [23, 24]. A conventional trigram model was built with Good-Turing discounting for bigrams and trigrams with counts lower than 7. No trigrams and no bigrams were discarded. The corpus was relatively small, so there were a lot of singletons with significant information. The language model perplexity of the test set was 48.

7.2. Automatically built classes

Models 4a and 4b use automatically built classes. Some experiments were performed to find the optimal number of classes. The number of classes was chosen to be 10, 100 and 1000, respectively. 9 experiments were performed to test all combinations (for classes of Slovenian and English words). 50 iterations of clustering algorithm were performed. The best results were obtained when English words were clustered into 10 classes, and Slovenian words into 100 classes. Some final classes showed slight semantic or morpho-syntactic resemblance between words, at least in English. In most classes words did not show any similarity. It was felt that data sparsity in the Slovenian part of the corpus was not reduced enough. The corpus was too small to build 'good' classes automatically. The Slovenian language (as a highly inflectional language) needs some additional linguistically oriented processing.

7.3. Model 4a

Model 4a is a conventional IBM Model 4. All word forms appear as unique tokens and were exposed as candidates for word-for-word alignments. Before training, the words were mapped into classes (as discussed in Section 7.2.).

After the training some interesting observations were made. Figure 5 shows the training-set and test-set perplexities computed after each iteration. Although the training-set perplexity continuously decreased, the test-set perplexity jumped at each transition point from one model to the next one (marked with the filled square in Figure 5). At the transition point, the final estimates of one model initialized the estimates of the next. 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 on those estimates obtained at the beginning of the training. The same observations have also been made in Czech-English experiments [18]. We speculated that the reason was the small size of the training corpus, so the translation probabilities become over-trained. Better

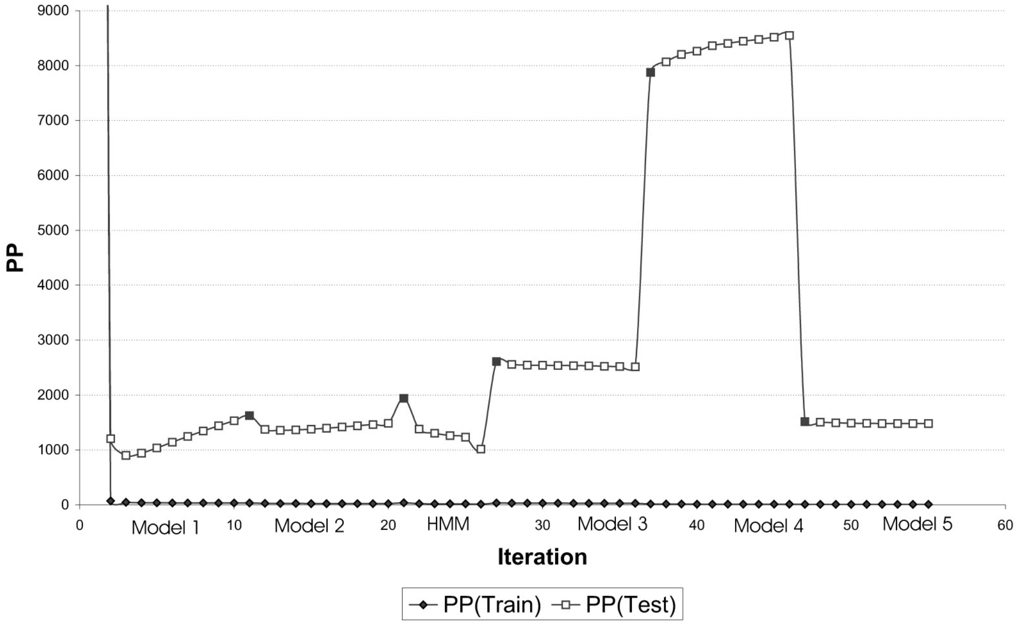

Fig. 5. Train-set and test-set perplexities during the Model 4a training.

alignments for training-sets did not lead to better translations of previously unseen test-sets. It was hoped the problem could be solved by reducing the sparsity of the training corpus.

Figure 6 shows the averaged distortion probabilities d_1 and $d_{>1}$. Relative displacement for one position has the highest probability $(d_{1,>1}(1|-,-) = 0.71)$. The reason is probably due to some determiners (the, a, an) and pronouns (I, we), which are zero-fertility words.

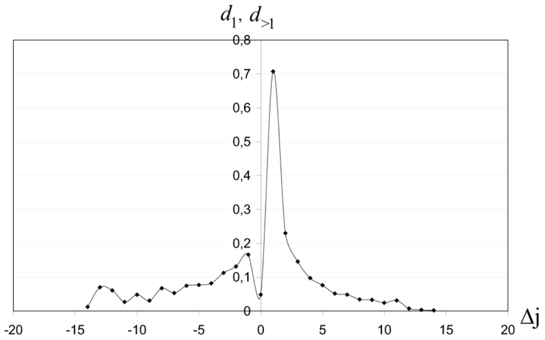

Fig. 6. Averaged distortion probabilities.

The probability p_1 was also determined in the training of Models 3, 4 and 5. Figure 7 shows the values of successive iterations (see the curve of Model 4a). The p_1 is relatively high until the end of Model 4 training.

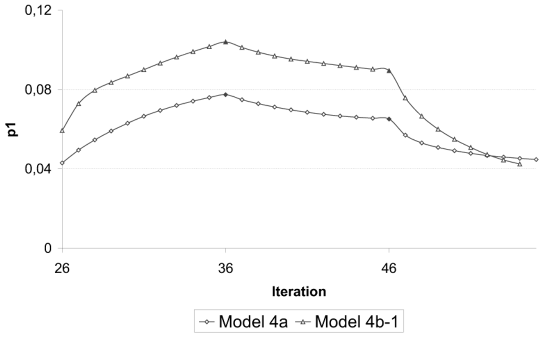

Fig. 7. Probability of NULL translation (p_1).

The evaluation of test-set decoding is given in the first rows of Tables 3 and 4. These values are used as a reference for further improvements.

7.4. Model 4b

Experiments with the lemmatised Slovenian part of the corpus were performed in two Models 4b. Lemmatising the Slovenian corpus reduced the data sparsity to a great extent (see Table 2). New clustering of Slovenian lemmas was performed. The Slovenian lemmas were automatically clustered into 100 classes. The training was repeated. Figure 8 shows the training-set

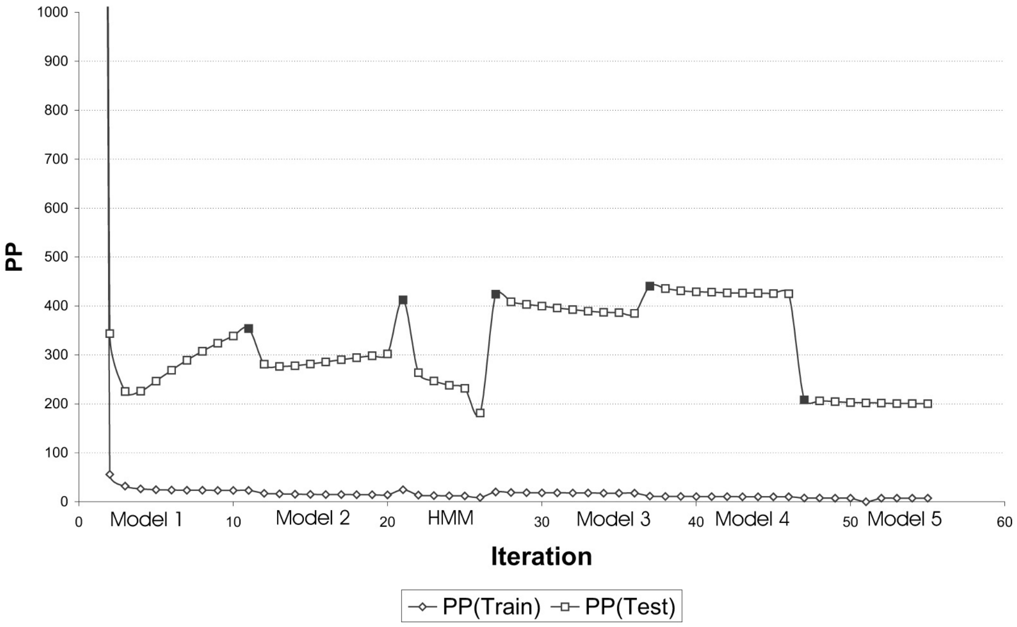

Fig. 8. Train-set and test-set perplexities during the Model 4b-1 training.

and test-set perplexities. It confirms the assumptions (from section 7.3) about Model 4 over-training. Each iteration of Model 4 produced a slight reduction of the test-set perplexity. Transition to Model 5 brought further improvements.

The test sentences were also lemmatised before decoding. The results of decoding are given in the second rows (see Model 4b-1) in Tables 3 and 4. The relative improvement of WER was 8% and of Bleu metric 21.71%.

In the Model 4b-2 we used our internal Slovenian-English dictionary, which we are building in our research laboratory. 10 000 dictionary entries were added, which occur in the vocabularies of our experiments. Note that the addition of a dictionary was possible, because the corpus was lemmatised. The results are given in the third rows (see Model 4b-2) in Tables 3 and 4. The WER and Bleu were improved in comparison to Model 4b-1. The relative improvement of WER was 1% and of Bleu metric 9%.

The drawback of Models 4b-1 and 4b-2 is the use of a lemmatiser during the on-line decoding process. An attempt was made to avoid this in Model 4c.

7.5. Model 4c

In Model 4c it was hoped to improve distortion probabilities. Words with the same morphological features were grouped into classes. The model is called Model 4c-1. Note that words were used as modelling units. Words were not lemmatised, because MSD codes are assigned to words and not to lemmas. The results are given in the fourth rows of Tables 3 and 4. The Model 4c-1 is better than Model 4a, but worse than Model 4b-1 and Model 4b-2.

Finally, in Model 4c-2, it was hoped to combine the advantages of Models 4b-2 and 4c-1. Translation probabilities were trained by Model 4b-2 and distortion probabilities by Model 4c-1. Translation probabilities, which had been associated with lemmas, were expanded to all word forms. The probabilities had to be normalized afterwards. The results are given in the last rows of Tables 3 and 4.

The final model (Model 4c-2) improves the WER of reference model (Model 4a) by 9% relatively and the value of Bleu metric by 39%. It can be seen that the results are very close to the results of Model 4b-2, without using a lemmatiser during the decoding. This model saves decoding time. It is particularly important, if requiring the translation system in real-time applications. Nevertheless, lemmas and MSD tags provide valuable information during the training of Model 4c-2.

Model	PP(Train)	PP(Test)
Model 4a	10	1324
Model 4b-1	7	229
Model 4b-2	11	121
Model 4c-1	11	1121

Table 3: Final translation model perplexities.

Model	WER (%)	Bleu (%)
Model 4a	69.3	18
Model 4b-1	63.7	22
Model 4b-2	62.9	24
Model 4c-1	67.9	20
Model 4c-2	63.0	25

Table 4: Evaluation of translation quality.

The training scheme of Model 4c-2 is our main achievement. Some (good and/or funny) translation examples obtained with our final Model 4c-2 are given in Table 5. Words are written with capital letters, because our goal is speech-to-speech translation (SST) system.

7.6. Comparison with related work

In this section we compare the results with related work.

Training corpus of the same size was used in experiments on Czech-English translation [4]. Using tectogrammatical parsing of Czech brought 95% relative improvement of Bleu metric. Using only lemmatisation has brought more than 100% relative improvement. In the latter

In:	IZDELKI ŽIVALSKEGA IZVORA KI NISO NAVEDENI IN NE ZAJETI NA DRUGEM MESTU
Ref:	PRODUCTS OF ANIMAL ORIGIN NOT ELSEWHERE SPECIFIED OR INCLUDED
Hyp:	ARTICLES OF ANIMAL ORIGIN ELSEWHERE SPECIFIED OR INCLUDED
In:	TA NOVA PARTICIJA MORA BITI ZDAJ PRIKLOPLJENA NEKAM V DREVO IMENIKOV
Ref:	NOW THIS NEW PARTITION MUST BE MOUNTED SOMEWHERE IN YOUR DIRECTORY TREE
Hyp:	THIS NEW PARTITION MUST BE NOW MOUNTED SOMEWHERE IN A DIRECTORY
In:	MORATA JE DEJALA ČE JE LE MOGOČE NAREDITI OTROKA
Ref:	THEY MUST SHE SAID PRODUCE A CHILD IF THEY COULD
Hyp:	BIDIRECTIONAL IF IT IS POSSIBLE EXTRACTING BABY
In:	GOTOVO JIM PREJ ALI SLEJ MORA PRITI NA MISEL DA JE TO TREBA STORITI
Ref:	SURELY SOONER OR LATER IT MUST OCCUR TO THEM TO DO IT
Hyp:	FALLEN THEM SOONER OR LATER IT MUST COME THOUGHT THAT IT MUST DO
In:	NAJHUJŠA PA JE BILA BOLEČINA V TREBUHU
Ref:	THE WORST THING WAS THE PAIN IN HIS BELLY
Hyp:	THE WORST BUT THERE WAS PAIN IN HIS BELLY

Table 5. Some examples. **In** denotes the sentence given as input to the decoder and **Hyp** is the output of the decoder. **Ref** is a reference sentence. The evaluation is done by comparing **Hyp** and **Ref**.

case they used the same translation as we did (GIZA++/ISI ReWrite Decoder). We speculate that the extent of improvement is due to the better type/token statistic of Czech WSJ training corpus (less singletons and zerotons) in comparison to IJS-ELAN corpus.

Polish-English translation includes grammar description and syntactic-semantic parsing. The translation system is transfer-based, so the focus is given to the process of converting an Oxford-PWN English-Polish dictionary to a format applicable for machine translation. It is difficult to compare the results, because the training data and the translation algorithm are not sufficiently described.

In experiments on Serbian-English SMT, Serbian words were reduced into stems. This procedure is, to some extent, comparable to lemmatisation. For translation direction Serbian to English stemming improves the translation by 8.3% (WER). Their results are comparable with ours. Our final model resulted in 9% relative improvement in WER.

8. Conclusion

In this paper, we have discussed different types of the IBM Model 4 translation model. The problem of data sparsity was outlined in experiments. Three new knowledge sources were added (lemmas, dictionary entries and MSD codes) successively, which partly reduced the sparsity. Finally, we have shown a method of overcoming the use of a lemmatiser in decoding (but not in training).

The results are far from being practically useful. One reason is the corpus size and its complexity. This problem could be solved. A more serious problem is the difficulty of the translation process. We hope to reduce it by adding more linguistic knowledge into the probabilistic framework.

Our work was done along the translation direction Slovenian to English. Translation in the opposite direction would require some changes in the system's architecture. At least the lemmatiser should be replaced with the module, which will produce the morphologically correct

Slovenian word forms from lemmas (using the context information). In addition, the language model for English should be replaced with the language model for Slovenian.

References

[1] Bleu,
URL: `http://www.ics.mq.edu.au/ szwarts/ Downloads.php`

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

[3] R. Brown, Example-based machine translation in the Pangloss system, *In Proceedings of COLING-96*, Copenhagen, Denmark, 1996.

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

[5] Peter Dirix, Vincent Vandeghinste, Ineke Schuurman, Techniques for a hybrid MT system, *In Proceedings of the meeting CLIN 2005*, Amsterdam, 2005.

[6] T. Erjavec, Compiling and Using the IJS-ELAN Parallel Corpus, *Informatica*, Vol. 26, 2002.

[7] T. Erjavec (ed.), Specifications and Notation for MULTEXT-East Lexicon Encoding. *MULTEXT-East Report*, Concede Edition D1.1F/Concede, Jožef Stefan Institute, Ljubljana. URL: `http://nl.ijs.si/ME/V2/msd/`

[8] T. Erjavec and S. Džeroski, Machine learning of morpho-syntactic structure: lemmatizing unknown Slovene words, *Appl. artif. intell.*, vol. 18, 2004.

[9] Evaluation Tools,
URL: `http://www.nist.gov/speech/tools/ index.htm`

[10] U. Germann, Greedy Decoding for Statistical Machine Translation in Almost Linear Time, *In Proceedings of the HLT-NAACL-2003*, Edmonton, AB, Canada.

[11] M. Jahr, Whittle – a corpus preparation tool, Egypt toolkit,
URL: `http://www.clsp.jhu.edu/ws99 /projects/mt/toolkit/`

[12] Krzysztof Jassem, Applying Oxford-PWN English-Polish dictionary to machine translation, *In Proceedings of the 9th EAMT Workshop*, 2004.

[13] Daniel Marcu and Ulrich Germann, The ISI ReWrite Decoder Release 1.0.0a,
URL: `http://www.isi.edu/licensed-sw/ rewrite-decoder/`

[14] M. S. Maučec, 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, 1997.

[15] S. Niessen, H. Ney, Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information, *Computational Linguistics*, 30(2), 2004.

[16] F. J. Och and H. Ney, A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, 29(1), 2003.

[17] F. J. Och, GIZA++: Training of statistical translation models,
URL: `http://www.fjoch.com/GIZA++.html`

[18] Y. All-Onaizan, J. Curin, M. Jar, K. Knight, J. Lafferty, I. D. Melamed, F. J. Och, D. Purdy, N. A. Smith and D. Yarowski, Statistical Machine Translation, JHU Workshop, Final report, Baltimore, Maryland, 1999.

[19] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation, *IBM research Report*, RC22176(W0109-022), 2001.

[20] Maja Popović, Slobodan Jovičić, Zoran Šarić, Statistical Machine Translation of Serbian-English, *In Proceedings of the SPECOM-2004*, St. Petersburg, Russia, 2004.

[21] M. Romih, P. Holozan, Slovensko-angleški prevajalni sistem (*Slovene - English translation system*), *In Proceedings of the conference Jezikovne tehnologije*, 2002.

[22] M. Romih, P. Holozan, Presis demo version, URL: `http://presis.amebis.si/prevajanje/`

[23] R. Rosenfeld, The CMU Statistical Language Modeling Toolkit, and its use in the 1994 ARPA CSR Evaluation, *In Proceedings ARPA SLT Workshop*, Austin, TX, 1995.

[24] R. Rosenfeld, The CMU Statistical Language Modeling Toolkit,
URL: `http://www.speech.cs.cmu.edu/SLM/ toolkit.html`

[25] D. Verdonik, M. Rojc, Z. Kačič, Creating Slovenian language resources for development of speech-to-speech translation components, In Proceedings of the conference LREC, Vol 4, 2004.

[26] Jernej Vičič and Tomaž Erjavec, Vsak začetek je težak : avtomatsko učenje prevajanja slovenščine v angleščino (*The beginning is always hard: training of machine translation from Slovene to English*), In Proceedings of the conference Jezikovne tehnologije, 2002.

[27] Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, Alex Waibel, The CMU Statistical Machine Translation System, *In Proceedings of the Machine Translation Summit IX*, New Orleans, Louisiana, USA, 2003.

Received: October, 2005
Revised: March, 2006
Accepted: April, 2006

Contact addresses:
Mirjam S. Maučec
Faculty of Electrical Engineering and Computer Science
University of Maribor
Smetanova 17, 2000 Maribor
Slovenia
e-mail: mirjam.sepesy@uni-mb.si

Zdravko Kačič
Faculty of Electrical Engineering and Computer Science
University of Maribor
Smetanova 17, 2000 Maribor
Slovenia

MIRJAM S. MAUČEC received her BSc and PhD degrees in computer science from the Faculty of Electrical Engineering and Computer Science at the University of Maribor in 1996 and 2001, respectively. She is currently a researcher at the same faculty. Her research interests include language modelling, statistical machine translation and computational linguistics.

ZDRAVKO KAČIČ is pressently a full professor at the Faculty of Electrical Engineering and Computer Science at the University of Maribor, and the Head of the Laboratory for digital signal processing. He was awarded his MSc degree in 1989 and PhD in 1992 at the same faculty. His research interests are analysis of complex sound scenes, systems for automatic speech recognition and creation of language resources.