Elektrotehniški vestnik 73(5): 273–278, 2006
Electrotechnical Review, Ljubljana, Slovenija

Statistical Alignment Models in Machine Translation from Slovenian to English

Mirjam Sepesy Maučec, Janez Brest, Zdravko Kačič

University of Maribor, Faculty of Electrical Engineering and Computer Science, Smetanova 17, Maribor, Slovenia (e-mail: mirjam.sepesy@uni-mb.si)

Abstract. This paper deals with statistical machine translation. The quality of translation system strongly depends on characteristics of the training corpus. In this paper we address the problem of very sparse training corpora. In languages with a very rich morphology, learning methods suffer from a significant sparseness problem. We present and compare various statistical models for computing word alignments that are the core of any translation model. The basic idea is to find a suitable training schedule for a specific task in terms of computing the "optimal" number of iterations for each translation model. For this purpose experiments are shown having been done on the IJS-ELAN bilingual corpus. We show that from the point of view of pure statistical translation the extent of data sparsity would most likely make such optimization impossible.

Key words: machine translation, word alignment, statistical model

Statistični modeli poravnave pri strojnem prevajanju iz slovenščine v angleščino

Povzetek. Članek obravnava problem statističnega strojnega prevajanja. Predstavimo in primerjamo različne uveljavljene statistične modele poravnav besed. Osredotočimo se na problem razpršenosti podatkov, ki je značilen za majhne korpuse. Osnovna ideja članka je izbira najprimernejše sheme učenja modela prevajanja v obliki iskanja "optimalnega" števila iteracij za posamezni model. Ideja je podkrepljena s poskusi, izvedenimi na korpusu IJS-ELAN. Pokažemo, da so s stališča strojnega prevajanja podatki preveč razpršeni, kar onemogoča učenje optimalnih parametrov. Poudariti velja, da smo se pri eksperimentih omejili le na uporabo stavčno poravnanih besed.

Ključne besede: strojno prevajanje, besedne poravnave, statistični model

1 Introduction

Machine Translation (MT) is widely considered among the most difficult tasks in natural language processing, and in artificial intelligence in general. A human translator can not possibly set down in a sufficient detail the "algorithm" he/she applies when translating a document. Instead, the idea was to make a system learn by itself how to translate. The growing availability of bilingual machine-readable text has stimulated interest in statistical methods for machine translation. Statistical methods require minimal human effort and can be created for any language pair that has enough training data.

In the machine translation community, the most widely studied language pairs are German-English [3], French-English and Japanese-English. 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 [7], Polish [8] and also last Slovenian [1, 2].

In this paper translation from Slovenian to English will be discussed. The focus is on the training schedule. It will be shown that the number of iterations of a specific alignment model is of a great importance.

1.1 Overview

In Section 2 the structure of statistical machine translation system is given. In Section 3 we first review the idea of general word-to-word alignment. Afterwards various statistical alignment models are described, sequentially used in our experiments. For each alignment model we give a set of probabilities it is composed of. Experiments are described in Section 4. A systematic empirical comparison of different training schemes is given. The optimal one is obtained with parameter tuning. The evaluation of the translation

Received 20 February 2006
Accepted 5 September 2006

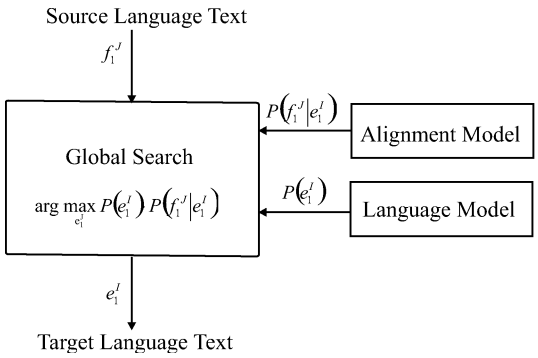

Figure 1. Architecture of statistical machine translation approach

quality was made with the automatically computable word-error-rate. Section 5 ends with a discussion of the achieved results.

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 the source string is a Slovenian sentence and the 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 posable target strings, the string with the highest probability as given by the Bayes' decision rule is chosen:

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

$P(e_1^I)$ 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. The architecture of the statistical translation approach is depicted in Figure 1. In this paper we focus on a translation model based on an alignment model.

3 Statistical Alignment Models

The training corpus is usually sentence-aligned and the idea is to perform word-alignment search. Very often, it is even difficult for a human to judge which words in a given target string correspond to which words in a source string. We can rewrite the probability $P(f_1^J|e_1^I)$ by introducing the "hidden" alignments $a_1^J = a_1...a_j...a_J$; $a_j \in 0, ..., I$:

$$P(f_1^J|e_1^I) = \sum_{a_1^J} P(f_1^J, a_1^J|e_1^I) \quad (2)$$

The different alignment models provide different decompositions of $P(f_1^J, a_1^J|e_1^I)$. The alignment a_1^J may contain alignments $a_j = 0$ with the "empty" word e_0 to account for source words that are not aligned with any target word.

Alignment models discussed in the paper are well known models from the literature [5]. They are trained in succession with the final parameter values of one model serving as the starting point for the next. The models will be presented in the same order.

3.1 Model 1

Model 1 uses the following formula:

$$P(f_1^J, a_1^J|e_1^I) = \frac{P(J|I)}{(I+1)^J} \prod_{j=1}^{J} P(f_j|e_{a_j}). \quad (3)$$

The alignment probability is composed of two probabilities: translation probability ($P(f_j|e_{a_j})$) and string length probability ($P(J|I)$). The word order does not affect the alignment probability. This model belongs to the group of models using zero order word alignment dependencies.

3.2 Model 2

The formula of Model 2 is:

$$P(f_1^J, a_1^J|e_1^I) = P(J|I) \prod_{j=1}^{J} P(a_j|j, I, J) \cdot P(f_j|e_{a_j}). \quad (4)$$

The alignment probability is composed of three probabilities: translation probability ($P(f_j|e_{a_j})$), string length probability ($P(J|I)$) and alignment probability ($P(a_j|j, I, J)$). The word order does affect the alignment probability. This model belongs to the group of models using zero order word-alignment dependencies.

3.3 HMM Model

The HMM-based model has the formula:

$$P(f_1^J, a_1^J|e_1^I) = \prod_{j=1}^{J} P(a_j|a_{j-1}, I) \cdot P(f_j|e_{a_j}). \quad (5)$$

The alignment probability is composed of two probabilities: translation probability ($P(f_j|e_{a_j})$) and alignment probability ($P(a_j|a_{j-1}, I)$). In the Hidden-Markov alignment model we assume a first-order dependence for the alignment a_j.

3.4 Model 3

Models 3 and 4 use an alternative alignment representation, the inverted alignments defining the mapping from the target to the source sentence (just the

opposite of Models 1, 2 and HMM Model). An important constraint for the inverted alignment is that all positions in the source language must be covered exactly once.

Model 3 is a zero-order alignment model like Model 2 including in addition fertility parameters. The fertility $\phi(e_i)$ of an English word e_i is the number of the Slovenian words aligned with it. The alignment probability is composed of four probabilities: translation probability ($P(f_j|e_i)$), fertility probability ($P(\phi_k|e_i)$), fertility probabilities for e_0 (p_0, p_1) and distortion probability ($P(j|i, J)$). Instead of fertilities $\phi(e_0)$ of the "empty" word, one single parameter $p_1 = 1 - p_0$ is used. It is the probability of putting a translation of "empty" word onto some position in a Slovenian sentence. Distortion probability gives the probability that an English word in position i is a translation of a Slovenian word in position j. The distortion probability depends on the length of the Slovenian sentence. The composition formula $P(f_1^J, a_1^J|e_1^I)$ is rather complicated. Readers interested in details are referred to paper [5].

3.5 Model 4

Model 4 is a first-order alignment model. The alignment position j of an English word depends on the alignment position of the previous English word j' (with non-zero fertility). Model 4 includes also fertilities. The alignment probability of Model 4 is composed of five probabilities: translation probability ($P(f_j|e_i)$), fertility probability ($P(\phi_k|e_i)$) fertility probabilities for e_0 (p_0, p_1) and two distortion probabilities: Distortion probabilities for the head word is denoted as $P_{=1}(\Delta j|\mathcal{A}(e_i), \mathcal{B}(f_j))$. Head word is the leftmost word of a set of source words aligned with the same target word. The remaining words are non-head words. Δj is the distance between the head of current translation, and the previous translation. It may be either positive or negative. Classes of words are used instead of words. $\mathcal{A}$ and $\mathcal{B}$ denote mapping of English and Slovenian words, respectively. Distortion probabilities for the non-head words is denoted as $P_{>1}(\Delta j|\mathcal{B}(f_j))$. In this case Δj denotes the distance between the head and non-head word.

Readers interested in details of the composition formula $P(f_1^J, a_1^J|e_1^I)$ are referred to paper [5].

3.6 Data sparsity

The problem of inflectionally rich languages is data sparsity. Each word-form is treated as distinct unit. Words, which are inflected forms of the same lemma, have "nothing in common" in alignment models. The extend of data sparsity will be apparent from experimental results. The reliability of probabilities estimated in each successive alignment model strongly depend on the data sparsity. in experiments we were looking for the best combination of alignment models owing to data sparsity.

Table 1. IJS-ELAN corpus

	Slo	Eng
Sentences	31,900	
Aver. sentence length	15.72	18.51
Words	498,906	587,481
- unique word	50,331	24,382
- singletons	24,830	10,575

4 IJS-ELAN Corpus

The experiments were performed on the IJS-ELAN corpus [1]. In the corpus there are parts with a Slovenian origin and an English translation, and parts with origins in English and translation in the Slovenian language. In spite of linguistic differences, we use all the parts in the same way. Half the corpus contains documents from the Slovenian government. The rest are two texts dealing with computers; one is about pharmaceuticals, and one is a literary work. All these collections are examples of a written language, except one, which contains speeches by the former President of Slovenia. 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). Although lemmas and MSD are a valuable source of information, in our experiments only words are used. The idea is to enlarge the corpus with bilingual texts for which the morphological analysis is not available. Some corpus statistics are collected in Table 1. It is interesting that the English part contains 18% more words than the Slovenian part.

The English sentence is on average three words longer than the 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 as 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). This data indicates the difficulty of the translation process.

5 Training, development and test sets

The corpus was partitioned into training, development and test sets. The training set is used to train

Table 2. Homogeneous corpus partitions

	Training set	Devel. set	Test set
Total sentences	12,064	1,539	1,539
Total words (Slo)	86,177	10,591	10,711
Total words (Eng)	97,258	11,929	12,016

Table 3. Training set statistic

	Slo	Eng
Words (sing.)	19,509 (58%)	11,454 (48%)
Co-occur. pairs	863,163	
- unique (sing.)	486,312 (79%)	

the probabilities. It is the largest one. The development set is used to optimize the parameters of the translation model. In our experiments the parameters will be the numbers of iterations for each model. The test set is used to evaluate the performance of the complete translation system, including the translation model, language model and search algorithm.

Long sentences significantly increase the computational complexity of training and decoding. We discarded sentences longer than 15 words. The division into training, development and test sets was done in ratio 8:1:1. The method in which a corpus is partitioned can significantly affect the experimental results. We used homogeneous data partitioning, in which development (test) sentences were taken from regular intervals through the corpus. The left-over data formed the training corpus. Characteristics of partition are given in Table 2.

The extend of the data sparseness problem is evident from Table 3 collecting statistics of the training set.

6 Initial translation model

The translation model was built using a program GIZA++ [5]. Models 1-4 were used as stepping stones. In the initial translation model ten iterations for training each model were performed.

The language model for the English language was made by using the CMU-SLM toolkit [4]. The whole English part of the IJS-ELAN corpus was used for training. A conventional trigram model was built with Good-Turing discounting.

Model 4 uses automatically built classes [9]. The English words were automatically clustered into 130 classes, and the Slovenian words into 1100 classes. The number of classes was set so as to be the same as the number of different MDS codes in the corpus. Ten iterations of the clustering algorithm were performed.

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

$$PP(Train) = 2^{-\frac{\sum_{(e,f)\in Train} \log P(f|e)}{N}} \qquad (6)$$

$Train$ denotes sentence-pairs from the training set. We measured the train-set perplexity ($PP(Train)$) and test-set perplexity ($PP(Test)$) (see Figure 2) of the translation model. The shapes of both curves are quite surprising. The first observation is that train-set perplexity increases when traversing from HMM model training to training Model 3. We speculate that modeling the relations between absolute positions in sentences (distortion probability) confuses the training. When Model 3 is eliminated from training, the final train-set perplexity of Model 4 increases slightly. Our conclusion is that other components of Model 3 positively influence Model 4 training and that distortion probability of Model 3 does not impact the distortion probabilities of Model 4. We decided to retain Model 3 training in our further experiments. More surprising is the shape of test-set perplexity. Despite the decreased train-set perplexity, test set perplexity continuously increases. The exception is HMM Model training that progressively decreases the test-set perplexity. The test-set perplexity comes to the lowest value in transition to Model 3, but this value was already obtained, when we just started to extract knowledge from the corpus.

With regard to test-set perplexities Models 3 and 4 can be eliminated from training. But they have to be kept because of the requirements of the decoder, which supports only Model 4. The decoding of test sentences was performed by an ISI ReWrite Decoder [6], The translation results were evaluated by the WER (word error rate) criteria, being the most strict among all evaluation metrics. It is computed as the minimum number of substitutions(Sub), insertions (Ins) and deletions (Del) that have to be performed to convert the hypothesis into the reference sentence. $Corr$ denotes correctly translated words. $S.Err$ are wrongly translated sentences.

Results of the decoding test set using the initial training model are given in the first row of Table 4 (T_{ini}).

Examining the initial results, the question arise as to what is the the optimal combination of all five models.

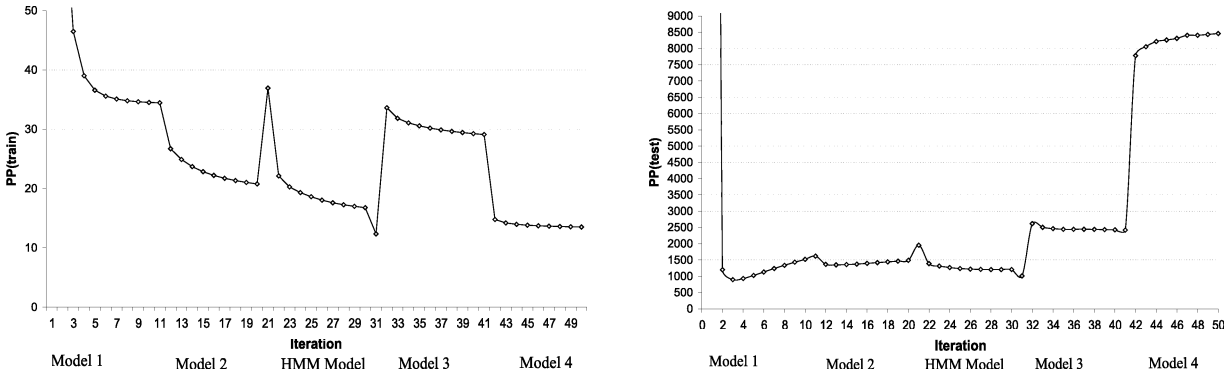

Figure 2. Train-set and test-set perplexities during training

Table 4. WER (in %) of decoding results

	Corr	*Sub*	*Del*	*Ins*	WER	S_{Err}
T_{ini}	37.3	39.9	22.9	8.5	71.2	89.4
D_{ini}	37.2	38.8	24.0	8.9	71.7	89.1
D_{fin}	38.0	38.6	23.3	9.0	71.0	88.1
T_{fin}	37.6	39.7	22.7	9.2	71.6	90.1

7 Parameters tuning

Before searching for optimal iteration numbers, the development set was analyzed. The results of the development set were almost the same as those of the test-set (see the second row in Table 4 (Dev_{ini})). The same curve was obtained also when the development-set perplexity was drawn. From this observation we concluded that both sets (test and development) share the same characteristics.

The goal of the translation system is to get correctly translated sentences. Because this is a hard nut to crack, we optimize the following sum:

$$f_{opt} = WER + S_{err}, \tag{7}$$

Parameters of the optimization function are: n_1, n_2, n_{HMM}, n_3 and n_4. n_1 denotes the number of iteration of Model 1. The explanation of other parameters is self-evident.

We performed 800 runs. During the optimization of f_{opt} value, each parameter was changed (from the parameter old value we generated its new one) using the following rule:

$$n_{i,new} = abs(n_{i,old} + N(0,\sigma)); \; i \in \{1,2,HMM,3,4\}, \tag{8}$$

where $N(0,\sigma)$ is Gaussian distribution. We set $\sigma = 10$ in our experiments. Also when a new parameter value was zero, we set it to 1.

In each run translation model training and development set decoding were made. Figure 3 shows how the numbers of iteration of each model changed during the optimization process. At the end values of the parameters were: $n_1 = 1$, $n_2 = 32$, $n_{HMM} = 53$, $n_3 = 12$ and $n_4 = 50$.

WER of the decoded development set was improved by 0.7% (absolute) and correctly translated sentences by 1% (absolute) (see the third row in Table 4). Decoding the test set with the tuned translation model gave slightly worse results (see the fourth row in Table 4). When comparing the shapes of perplexity curves (see Figure 4), it was noted that they match almost in each point. Both sets share the same relation in terms of data sparsity with train-set, but at the same time they have "nothing in common". Based on this two observations our conclusion is that the possibility of predicting an unseen text with the probability learned from the training set is scarce.

8 Conclusion

The discussed statistical alignment models would be of a true value if data sparsity of the corpus were reduced. This will be done by enlarging the parallel corpus. We plan to proceed our research on much larger SVEZ-IJS corpus. In pursuance of this target, lemmas and MSD labels will be included in translation model training in our future experiments.

9 References

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

[2] J. Vičič, "Statistično strojno prevajanje naravnih jezikov", *Master thesis*, 2002.

[3] W. Wahlster, *"Verbmobil: Foundations of Speech-to-Speech Translation"*, Springer, 2002.

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

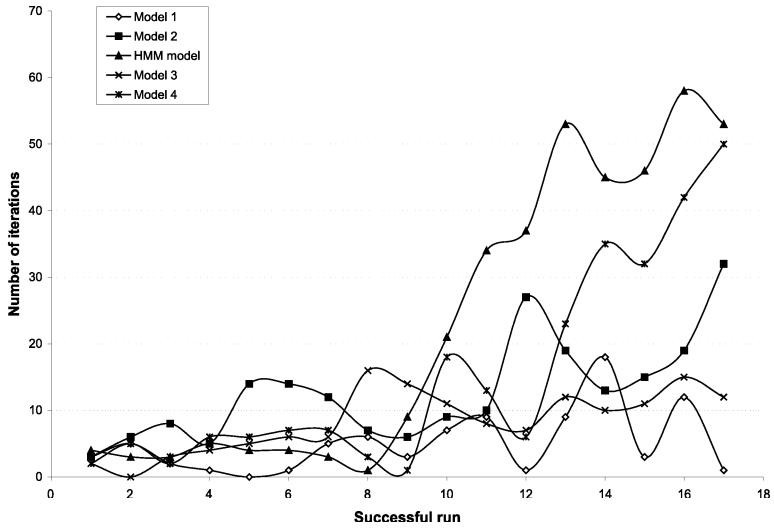

Figure 3. Number of iterations for each model in succesful runs (runs improving f_{opt} are depicted)

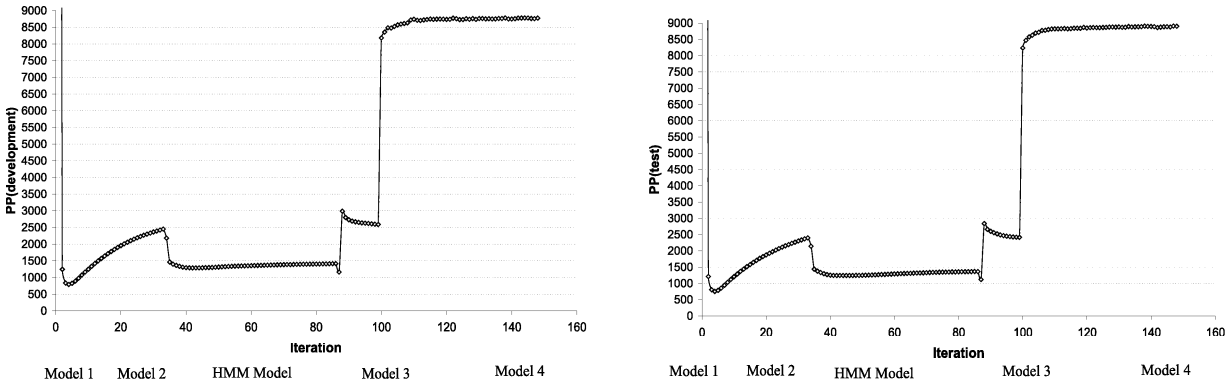

Figure 4. Development-set and test-set perplexities during training

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

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

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

[8] Krzysztof Jassem, "Applying Oxford-PWN English-Polish dictionary to machine translation", *9th EAMT Workshop*, 2004.

[9] M. S. Maučec, J. Brest, Z. Kačič, V. Žumer, "Statistično modeliranje naravnega jezika pri avtomatskem razpoznavanju govora z uporabo heterogenega računalniškega sistema.", *Elektroteh. vestn.*, 2000, letn. 67, t. 1, str. 55-61.

Mirjam S. Maučec received her B.Sc. and Ph.D. degrees in computer science of the Faculty of Electrical Engineering and Computer Science at the University of Maribor, Slovenia, 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.

Janez Brest received his B.Sc., M.Sc. and Ph.D. degrees in 1995, 1998 and 2001 respectively, from the Faculty of Electrical Engineering and Computer Science of the University of Maribor, Slovenia. In 1993 he joined the Laboratory for Computer Architecture and Programming Languages. He is currently an Assistant Professor at the the same faculty. His research includes evolutionary computing, optimization, programming languages and machine learning.

Zdravko Kačič received his M.Sc. and Ph.D. degrees in 1998 and 2001 respectively, from the Faculty of Electrical Engineering and Computer Science of the University of Maribor, Slovenia. He is currently Full Professor and the Head of the Laboratory for digital signal processing at the same faculty. His research interests are in systems for automatic speech recognition, spoken dialog systems and multimodal communication.