Training Corpus Size and Statistical Machine Translation Quality

Nicola Ueffing and Hermann Ney
{ueffing,ney}@cs.rwth-aachen.de

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology, D-52056 Aachen, Germany

Supervisor:	Prof. Dr.-Ing. H. Ney
Type:	PhD thesis
GI subjects:	Natural Language Processing (1.3.1), Machine Learning (1.1.3)

Abstract
Statistical machine translation systems rely heavily on the amount and on the type of data that is available for the training of the system. We systematically investigate the effect of the size of the bilingual corpus for the training of the translation model on translation quality. We further show the system can be improved using only a monolingual corpus. Results are presented on two translation tasks.

1 Introduction

The goal of statistical machine translation is the translation of a string $f_1^J = f_1 \dots f_j \dots f_J$ in the (natural) source language into a string $e_1^I = e_1 \dots e_i \dots e_I$ in the (natural) target language. This target string is chosen according to Bayes' decision rule

$$\begin{aligned}\hat{e}_1^I &= \operatorname*{argmax}_{e_1^I} \left\{ Pr(e_1^I | f_1^J) \right\} \\ &= \operatorname*{argmax}_{e_1^I} \left\{ Pr(f_1^J | e_1^I) \cdot Pr(e_1^I) \right\}\end{aligned}$$

A statistical machine translation system consists of three components: the **translation model**, $Pr(f_1^J|e_1^I)$, the **language model** of the target language, $Pr(e_1^I)$, and the **search**, denoted by the argmax operation. Among all possible target strings, we have to find the target string that maximizes the probability calculated from the models. Those probabilities are estimated on corpora: for the translation model, a bilingual corpus containing translations of the same text in the source and the target language is needed; and the language model is trained on a monolingual corpus in the target language. Usually, this monolingual corpus is the target language part of the bilingual corpus, but – as we will see later – enlarging only this monolingual corpus yields a gain in translation quality.
Data collection – especially of the bilingual corpus – is a very expensive task in the construction of a statistical machine translation system. Therefore, an investigation on the amount of data required to ensure sufficient quality is of great interest for the compilation of new corpora.

2 Experimental Setup and Evaluation Criteria

The translation system we used employs the translation model IBM-4 from the pioneering work of the IBM research group [1]. It is based on dependencies between single words in the source and the target language string (in contrast to phrase-based or syntax-based translation models [2, 3, 4, 5, 6]). A description of our translation system can be found in [7, 8]. The language model is a trigram model, i.e. it predicts the next word on the basis of the two preceding words.
The quality of the output of our machine translation system is measured automatically by comparing the produced translation to a given reference translation (produced by a human translator). The two following criteria are used: for evaluation

- WER (**w**ord **e**rror **r**ate):
  The word error rate is based on the Levenshtein distance. It is computed as the minimum number of substitution, insertion and deletion operations that have to be performed to convert the generated string into the reference string.
- BLEU (**bil**ingual **e**valuation **u**nderstudy):
  [9] have proposed a method of automatic machine translation evaluation, which they call "BLEU". It is based on the notion of modified n-gram precision. In contrast to the word error rate, BLEU conveys quality (i.e. higher scores are better).

Table 1: Verbmobil Training and Test Corpus Statistics.

		German	English
Training	Sentences	58 073	
	Words	519 523	549 921
	Words without Punctuation Marks	418 979	453 632
	Average Sentence Length	8.9	9.5
Vocabulary	Size	7 911	4 648
Test	Sentences	251	
	Words	2 627	2 866
	Average Sentence Length	10.5	11.4

3 The Size of the Bilingual Corpus

We performed experiments on the German-English Verbmobil corpus [10]. It consists of spontaneously spoken dialogues in the appointment scheduling domain. German source sentences were translated into English. The training and test corpus statistics are given in Table 1. The translation direction is from German into English.

In order to investigate the effect of the amount of bilingual data on translation quality, we ran the translation model training procedure with 10% to 100% of the bilingual corpus, increasing in steps of 10%. For the training of the language model, we used two setups: First, we estimated the parameters of the model on the target language part of the corpus that was used for the translation model training, i.e. the language model corpus size was gradually increased as well. Secondly, we used the whole 58 073 sentences for the training of the language model – independently of the amount of data that was used for the translation model training. We then carried out translation experiments with those different systems. In Figure 1a, we see the results of the translations in the first setup. The word error rate decreases significantly for growing corpus sizes – from 46.92% down to 40.93% – and the (quality measure) BLEU increases steadily from 43.00% to 50.54%. If we train the language model on all the available data, the saturation point of the error rate/BLEU score is reached much earlier: already 30% of the bilingual training corpus lead to the same quality of the results as the whole training corpus (see Figure 1b). This fact inspired us to increase the monolingual corpus for language model training as described in the next section.

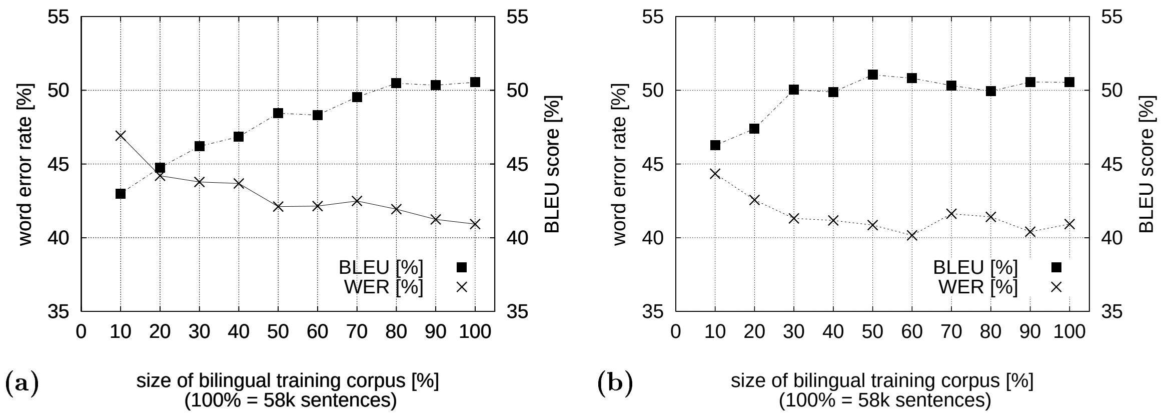

Figure 1: Word Error Rate/BLEU Score for Different Sizes of the Translation Model Training Corpus, Using (a) the Same Corpus Size (b) the 58 073 Sentences for Language Model Training.

4 The Size of the Monolingual Corpus

The compilation of bilingual corpora is much more costly than that of monolingual corpora which are widely available for many domains. Thus, we gain a lot if we can use monolingual resources to improve machine translation quality. The experiments in the last section indicate that ameliorating translation performance by enlarging the language model training corpus is possible .

If we want to increase the amount of data only for one of the languages involved (in our case: the

Table 2: Canadian Hansards Training and Test Corpus Statistics.

		French	English
Training	Sentences	49 000	
	Words	743 903	816 964
	Words without Punctuation Marks	664 058	730 880
	Average Sentence Length	15.2	16.7
Vocabulary	Size	19 831	24 892
Test	Sentences	500	
	Words	8 749	7 946
	Average Sentence Length	17.5	15.9

target language), we have to keep in mind that it has to be from the same domain as the bilingual corpus and it has to cover the same vocabulary. Therefore, we chose the Canadian Hansards for our experiments (translating French into English): We have 1.5 million sentences available such that we can easily extend the English corpus within the domain and vocabulary to investigate the effect of this. The Hansards consist of the proceedings of the Canadian parliament which are kept both in French and in English. The bilingual training corpus comprises 49 000 sentences in each language (see Table 2). From 1.5 million English sentences of the Hansards corpus, we filtered those that have the same vocabulary as our training set which yielded 1 272 131 additional sentences. Again, we subjoined increasing amounts of data to our training material: 0% to 100% of the supplementary sentences were added to the target language corpus in steps of 10%.

This successively improves translation quality as Figure 2a shows: the error rate decreases from 71.89% to 67.98%, and the BLEU score climbs from 18.24% up to 24.79%. The translation model parameters were not altered, but obviously a better language model enables the system to produce better translations. This is not surprising as one weakness of statistical machine translation systems is the ungrammaticality of their output.

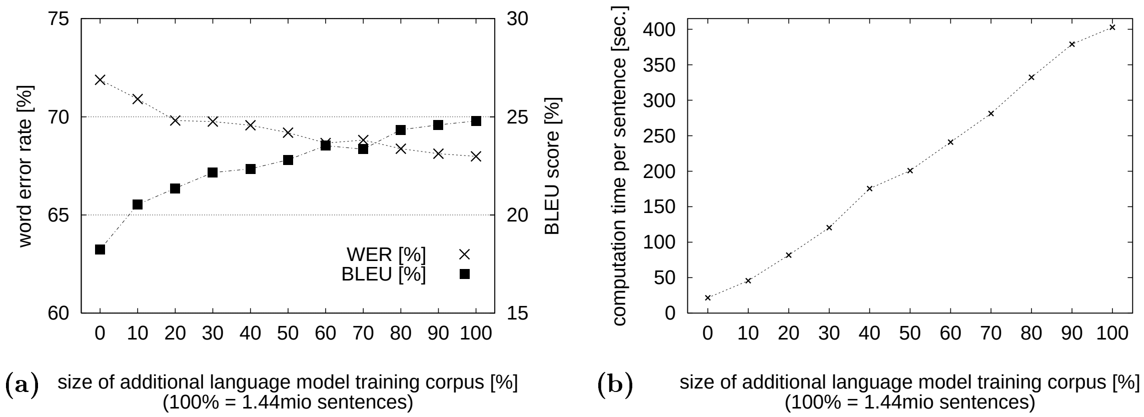

Figure 2: Word Error Rate/BLEU Score and Average Computation Time per Sentence as a Function of Language Model Training Corpus Size

As can be seen in Figure 2b, the extension of the language model size also increases computation time. The system trained on the smallest corpus needs 22 seconds on an average for the translation of one sentence, and the time consumption grows almost linearly up to 400 seconds for the best system. This is due to the fact that much more language model events have to be considered during search. Table 3 indicates the average number of predecessor events for each word in the vocabulary. This number rises significantly the larger the language model corpus gets: in the smallest language model, 20.87 predecessors per word occur on average, whereas in the largest language model, there are 250 events on average.

Table 3: Average Number of Predecessor Events per Word in the Language Model.

Additional Data [%]	0	10	20	30	40	50	60	70	80	90	100
# Predecessors	20.87	55.41	83.66	108.81	132.09	153.94	174.77	194.52	213.56	231.81	249.23

5 Conclusion

We presented a systematic investigation on the effect of the training corpus size on translation quality. The outcome of a variation in the amount of bilingual as well as monolingual data were studied in our experiments. We found that already an extension of the monolingual corpus for the target language yields an improvement in translation quality. This point is particularly interesting for the compilation of new corpora for statistical machine translation as monolingual resources are much easier to obtain.

6 Acknowledgement

This work was partly supported by the LC-STAR project by the European Community (IST project ref. no. 2001-32216).

References

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

[2] D. Marcu and W. Wong. A Phrase-Based, Joint Probability Model for Statistical Machine Translation. In *Proc. Conf. on Empirical Methods for Natural Language Processing*, Philadelphia, PA, pages 133–139, July 2002.

[3] F. J. Och. *Statistical Machine Translation: From Single-Word Models to Alignment Templates.* PhD thesis, Computer Science Department, RWTH Aachen – University of Technology, Germany, 2002.

[4] D. Wu and H. Wong. Machine Translation with a Stochastic Grammatical Channel. In *COLING-ACL '98: 36th Annual Meeting of the Assoc. for Computational Linguistics and 17th Int. Conf. on Computational Linguistics*, Montréal, Québec, Canada, pages 1408–1414, August 1998.

[5] K. Yamada and K. Knight. A Syntax-Based Statistical Translation Model. In *Proc. 39th Annual Meeting of the Assoc. for Computational Linguistics - joint with EACL*, Toulouse, France, pages 523–530, July 2001.

[6] R. Zens, F. J. Och, and H. Ney. Phrase-Based Statistical Machine Translation. In *Proc. German Conf. on Artificial Intelligence (KI 2002)*. Springer Verlag, September 2002.

[7] C. Tillmann and H. Ney. Word re-ordering and DP beam search for statistical machine translation. *to appear in Computational Linguistics*, 2002.

[8] N. Ueffing, F. J. Och, and H. Ney. Generation of Word Graphs in Statistical Machine Translation. In *Proc. Conf. on Empirical Methods for Natural Language Processing*, Philadelphia, PA, pages 156–163, July 2002.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, September 2001.

[10] W. Wahlster, editor. *Verbmobil: Foundations of speech-to-speech translations.* Springer-Verlag, Berlin, 2000.