Using a Bilingual Context in Word-Based Statistical Machine Translation

Christoph Schmidt, David Vilar and Hermann Ney

Chair of Computer Science 6
RWTH Aachen University
{schmidt,vilar,ney}@cs.rwth-aachen.de

Abstract. In statistical machine translation, phrase-based translation (PBT) models lead to a significantly better translation quality over single-word-based (SWB) models. PBT models translate whole phrases, thus considering the context in which a word occurs. In this work, we propose a model which further extends this context beyond phrase boundaries. The model is compared to a PBT model on the IWSLT 2007 corpus. To profit from the respective advantages of both models, we use a model combination, which results in an improvement in translation quality on the examined corpus.

1 Introduction

The goal of machine translation is to translate a text from one natural language into another using a computer. In statistical machine translation, the process of translating is modelled as a statistical decision process.

The IBM-models proposed in the early 1990s were single-word-based models [1]. A characteristic of the single-word-based approach is that lexicon probabilities are modelled for single words. Consequently, the context in which a word is used does not influence its translation probability.

The phrase-based approach tries to overcome this disadvantage by learning the translation of whole phrases instead of single words. Here, "phrase" is not used in the linguistic sense but simply refers to a sequence of words. (A more detailed description of the phrase-based approach can be found in [2].) As the phrase-based approach translates phrases independently, words outside the phrase are not considered for its translation. Moreover, the PBT model makes some independence assumptions which seem to be arbitrary, e.g. the assumption that all segmentations of the source sentence into phrases are equally likely.

To overcome these deficiencies, the model proposed in this work considers a bilingual context beyond phrase boundaries. This approach is similar to the N-gram model presented in [3]. However, the model presented here remains at the target word level. As the model conditions its translation on both the words of the source and the target sentence, we will refer to it simply as the "conditional model".

The remaining part of this work is organized as follows: in the next section, we will briefly sketch the basics of statistical machine translation and describe the log-linear approach. In Section 3, the conditional model is introduced. Section 4 and Section 5 describe the search process and the feature functions used in the log-linear approach.

In Section 6, we propose a model combination of the PBT model and the conditional model. Section 7 discusses the experimental results obtained on the IWSLT 2007 corpus. The last section gives a conclusion and an outlook for possible future research.

2 Statistical Machine Translation

In statistical machine translation, a given source language sentence $f_1^J = f_1 \ldots f_J$ has to be translated into a target language sentence $e_1^I = e_1 \ldots e_I$. According to Bayes' decision rule, to minimize the sentence error rate we have to choose the sentence $\hat{e}_1^I$ which maximizes the posterior probability[1]:

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \{Pr(e_1^I|f_1^J)\} . \tag{1}$$

The posterior probability $Pr(e_1^I|f_1^J)$ is modelled directly using a log-linear model [4]:

$$p(e_1^I|f_1^J) = \frac{\exp(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J))}{\sum_{\tilde{e}} \exp(\sum_{m=1}^M \lambda_m h_m(\tilde{e}, f_1^J))} . \tag{2}$$

Inserting Equation (2) into Bayes' decision rule (1) and simplifying, we obtain:

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \{Pr(e_1^I|f_1^J)\} \tag{3}$$

$$= \underset{e_1^I}{\operatorname{argmax}} \{\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\} . \tag{4}$$

In the log-linear model, different knowledge sources can be easily combined using *feature functions* $h_m(e_1^I, f_1^J)$. Statistical translation systems typically use bilingual features such as translation probabilities and monolingual features such as the target language model. The conditional model which models the posterior probability $Pr(e_1^I|f_1^J)$ can also be used as one feature function in the log-linear approach.

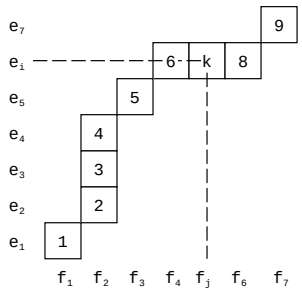
The model scaling factors λ_m are estimated on a development corpus to optimize some performance measure, usually BLEU.

3 The Conditional Model

3.1 Motivation

The main improvement of the phrase-based translation (PBT) model over the single-word-based (SWB) translation model is the extension of the context taken into account when translating a word: While the SWB model translates words individually and uses context information only in the target language model, the PBT model translates whole sequences of words. Nonetheless, the PBT model does not take into account words outside the phrase to be translated. In the following, the conditional model which considers a context beyond phrase boundaries is developed.

[1] Note that in this work, Pr is used to indicate true probabilities, while p denotes probability models.



Relations between i,j and k in the figure:

- $k = 7$
- $\mathcal{A}_k = (5, 6)$
- $i(k) = 6$
- $j(k) = 5$
- $k_{i=6} = 8$
- $k_{j=2} = 4$

Fig. 1. Relations between the alignment indices i,j and k

3.2 The Conditional Model

A word *alignment* $\mathcal{A}$ is a relation $\mathcal{A} \subseteq J \times I$ such that $(j, i) \in \mathcal{A}$ if word f_j in the source sentence corresponds to word e_i in the target sentence. In general, alignments allow for many-to-many relations between source and target words.

The conditional model is monotone: it translates words in the same order in which they appear in the source sentence. However, the word order of the source language often differs from that of the target language. To overcome this problem, the source sentence is first reordered in such a way that the reordered sentence can then be translated monotonously. In this section, a reordered source sentence and a monotone alignment are presumed. Section 3.3 will explain how to obtain such a reordering during training.

Alignment points (j, i) are numbered by an index k:

$$\mathcal{A} : \{1, \ldots, K\} \rightarrow \{1, \ldots, J\} \times \{1, \ldots, I\} \tag{5}$$

$$\mathcal{A}_k \; = \; (j(k), i(k)) \tag{6}$$

This indexing is done consecutively from $\mathcal{A}_1 = (1, 1)$ to $\mathcal{A}_K = (J, I)$ (see Fig. 1).

A source word can be aligned to several target words and vice versa. The functions k_i and k_j obtain the alignment point with the highest index for a given source/target word:

$$k_i = \max_{k \,:\, \exists j \,:\, \mathcal{A}_k = (j,i) \in \mathcal{A}} k \qquad , \qquad k_j = \max_{k \,:\, \exists i \,:\, \mathcal{A}_k = (j,i) \in \mathcal{A}} k$$

Starting from the posterior probability $Pr(e_1^I | f_1^J)$, the alignment is introduced as a hidden variable $\mathcal{A}$:

$$Pr(e_1^I | f_1^J) = \sum_{\mathcal{A}} Pr(e_1^I, \mathcal{A} | f_1^J) \tag{7}$$

$$= \sum_{\mathcal{A}} \prod_{i=1}^{I} Pr(e_i | e_1^{i-1}, f_1^J, \mathcal{A}) Pr(\mathcal{A}_{k_i} | \mathcal{A}_1^{k_{i-1}}, f_1^J) \tag{8}$$

In Equation (8), the probabilities are decomposed using the chain rule $Pr(x_1^N) = \prod_n Pr(x_n|x_1^{n-1})$. Furthermore, the joint probability of target words and alignment is separated.

The conditional model restricts the dependency of a target word to the source and target words which are aligned by the last m alignment points:

$$Pr(e_i|e_1^{i-1}, f_1^J, \mathcal{A}) = p(e_i|e_{i(k_i-m)}^{i-1}, f_{j(k_i-m)}^{j(k_i)}, \mathcal{A})\,. \tag{9}$$

This Markov assumption of order m is similar to that of n-gram models in language modelling: n-gram models restrict the dependency of a word to its $n-1$ predecessors. m is called *model order*.

Additionally, the dependency of the alignment probability $Pr(\mathcal{A}_{k_i}|\mathcal{A}_1^{k_{i-1}}, f_1^J)$ is restricted to the previous alignment point $\mathcal{A}_{k_{i-1}}$ and the current source word $f_{j(k_i)}$. As only the differences between $j(k_i)$ and $j(k_{i-1})$ are considered, $\mathcal{A}_{k_i} = (j(k_i), i)$ can be simplified to $j(k_i)$:

$$Pr(\mathcal{A}_{k_i}|\mathcal{A}_1^{k_{i-1}}, f_1^J) = p(j(k_i)|j(k_{i-1}), f_{j(k_i)})\,. \tag{10}$$

In a second step, the conditional probability $p(j(k_i)|j(k_{i-1}))$ is replaced by the Bakis model known from speech recognition [5]:

$$\delta = \begin{cases} 0 & \text{if } j(k_i) = j(k_{i-1}) \\ 1 & \text{if } j(k_i) = j(k_{i-1}) + 1 \\ 2 & \text{if } j(k_i) > j(k_{i-1}) + 1 \end{cases} \tag{11}$$

Fig. 2 shows the different values δ can take. In this example, the target word to be translated is e_6. Consequently, k_6 is 8. For a model order of $m = 5$, the gray alignment points are the considered context, which corresponds to the words e_3^5 and f_2^6.

Applying the assumptions of Equations (9) and (10) leads to:

$$Pr(e_1^I|f_1^J) = \sum_{\mathcal{A}} \prod_{i=1}^{I} p(e_i|e_{i(k_i-m)}^{i-1}, f_{j(k_i-m)}^{j(k_i)}, \mathcal{A}) \cdot p(j(k_i)|j(k_{i-1}), f_{j(k_i)}) \tag{12}$$

$$= \sum_{\mathcal{A}} \prod_{i=1}^{I} p(e_i|\delta, e_{i(k_i-m)}^{i-1}, f_{j(k_i-m)}^{j(k_i)}) \cdot p(\delta|f_{j(k_i)}) \tag{13}$$

The alignment information is contained in the function k_i, and consequently $\mathcal{A}$ is omitted in (13).

Instead of summing over all possible alignments, the maximum-approximation which considers only the alignment leading to the highest value is used. This assumption reduces the complexity of the search procedure.

$$Pr(e_1^I|f_1^J) \approx \max_{\mathcal{A}} \prod_{i=1}^{I} p(e_i|\delta, e_{i(k_i-m)}^{i-1}, f_{j(k_i-m)}^{j(k_i)}) \cdot p(\delta|f_{j(k_i)}) \tag{14}$$

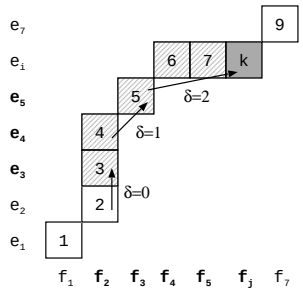


Fig. 2. Delta-notation, alignment points taken into account for m=5

The probabilities of the model are estimated from the training corpus via relative frequencies. With high model orders, the probabilities cannot be estimated reliably. To solve this data sparseness problem, the probabilities are smoothed using linear discounting [6]. Whenever a context of order m is encountered which has not been seen during training, shorter contexts are considered instead.

3.3 Monotonization of Alignments

Since the conditional model is defined to be monotone, the training data has to be transformed to obtain reordered source sentences and monotone alignments. Given an unaligned parallel training corpus, first general alignments are computed under the following restrictions:

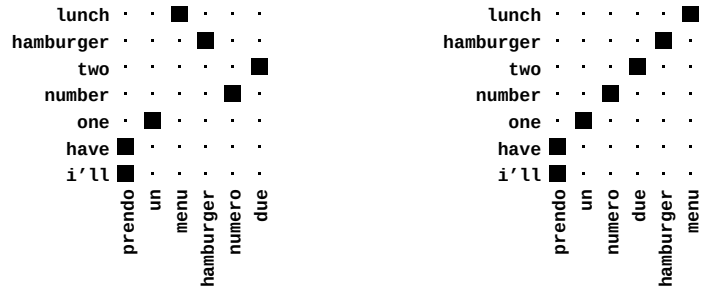
1. Each source and target word has to be aligned to at least one word. The resulting alignments are therefore called *full-coverage alignments*.
2. If a source word is aligned to several target words, these words have to be consecutive.
3. If a source word is aligned to several target words, these target words cannot be aligned to other source words.

Alignments which fulfill the above conditions can then be transformed into monotonic alignments by reordering the source words appropriately. It is sufficient to sort the alignment columns according to the smallest row index of the alignment points they contain. For a detailed description cf. [7]. Fig. 3 gives an example of such a reordering. $\breve{f}_1^J$ indicates the reordered source sentence, $\breve{\mathcal{A}}$ the monotone alignment.

4 The Search Process

The goal of the translation process is to obtain the target sentence which maximizes the translation probability $Pr(e_1^I|f_1^J)$. A beam-search algorithm is used.

To allow for a difference in word order, the source sentence is reordered using reordering graphs. These generate a restricted subset of permuted source sentences [2]. A simple cost function $reorder(f_1^J, \breve{f}_1^J)$ which applies higher costs to non-monotonic translations is used as a feature function. In the experiments, reordering graphs with



(a): Original alignment, source sentence f

(b): Monotone alignment $\breve{\mathcal{A}}$, reordered source $\breve{f}$

Fig. 3. A monotone alignment obtained by reordering the source sentence f

local constraints yielded best results. With these constraints, only neighboring words within a certain window are permuted.

5 Feature Functions

In the log-linear approach, six feature functions were used:

$h_1(e_1^I, f_1^J) = log\left(Pr(e_1^I \mid f_1^J)\right)$	posterior probability
$h_2(e_1^I, f_1^J) = log\left(Pr(f_1^J \mid e_1^I)\right)$	inverse posterior probability
$h_3(e_1^I, f_1^J) = I$	word penalty
$h_4(e_1^I, f_1^J) = log\left(Pr(e_1^I)\right)$	target language model
$h_5(e_1^I, f_1^J) = log\left(Pr(\breve{f}_1^J)\right)$	language model for reordered source
$h_6(e_1^I, f_1^J) = reorder(f_1^J, \breve{f}_1^J)$	reordering cost

Often, the translation system produces very short sentences e_1^I, because for each additional word, the translation probability gets smaller. The word penalty feature function can counterbalance this tendency. With a negative weight λ_3, a bonus is added for each additional word. The language model for $Pr(\breve{f}_1^J)$ is trained on the reordered source sentences of the training corpus. It is able to learn reordered word sequences which occur in the training corpus and complements the simple reordering cost h_6.

6 Combining the PBT Model and the Conditional Model

Additionally, the PBT model and the conditional model can be combined to take advantage of their respective benefits:

- The PBT model is able to reorder whole phrases instead of single words. It can rule out many reorderings in which related words are separated.
- The conditional model considers a context beyond phrase boundaries.

A standard method for combining different models is N-best list rescoring: First, one model is used to translate the source text. For each sentence, the N best translations are stored along with their model costs. In a second step, the other model is used to score the translated sentences of the first model. Given a translation made by the first model, the cost of this translation with respect to the second model is calculated. In the end, a log-linear combination of both costs is calculated. The weights of this combination are optimized on the development corpus.

In the model combination, the reordering which was computed by the PBT model when translating a sentence is also used by the conditional model. The conditional model performs a monotone translation of the reordered source sentence produced by the PBT model. Thus, the model combination takes advantage of the better reordering capabilities of the PBT model. The model combination led to an improvement over the PBT model, as can be seen in the following section.

7 Translation Results

7.1 Evaluation Criteria

Evaluating the quality of a translation is in itself a difficult task. In the experiments presented here, we rely on two criteria, TER and BLEU.

- TER (translation edit rate) [8] : The TER criterion is a recent refinement of the WER criterion. The WER criterion is defined as the edit distance (minimum number of insertions, deletions and substitutions) between the translation and a reference translation. In addition to this, the TER allows for a sequence of contiguous words to be moved to another place. This enhancement is natural, as often phrases can be placed at different position of the sentence without altering its meaning.
- BLEU (bi-lingual evaluation understudy) [9] : BLEU is the current de facto standard in machine translation evaluation. When using several translation references, BLEU can capture the variability translations can have. It evaluates the translated sentence by calculating an n-gram precision for $n \in \{1, 2, 3, 4\}$. In addition, a brevity penalty is calculated to penalise translations which are too short. Note that a good translation is indicated by a high BLEU score. BLEU is said to have a high correlation with human evaluation. The parameter of the model presented in this work were optimized with respect to BLEU.

7.2 Corpus Statistics

For the following experiments, the two language pairs Chinese-English and Italian-English were chosen from the corpus used in the "International Workshop on Spoken Language Translation" IWSLT 2007 [11]. The corpus is a further development of the the multilingual BTEC ("Basic Travel Expression Corpus") corpus which contains typical phrases and sentences from the travelling domain. In Table 1, some corpus statistics are summarized. In the Chinese case, six reference translations were given for each sentence in the test set, in the Italian case, four reference translations were given.

Table 1. Corpus Statistics IWSLT 2007 Chinese-English and Italian-English

		Chinese	English	Italian	English
Training data:	Sentences	42,942		22,995	
	Running Words	390,335	420,431	164,715	222,005
	Vocabulary	10,385	9,933	10,329	7,794
	Singletons	3,696	3,937	4,729	3,355
Test data:	Sentences	489	$6 \cdot 489$	724	$4 \cdot 724$
	Running Words	3,256	22,574	6,540	36,725
	Vocabulary	885	1,527	735	940
	OOVs (running words)	70	4,377	449	6,799
	OOVs (in vocabulary)	69	394	110	288

To see whether the extension of the context beyond phrase boundaries leads to an improvement in translation quality, the conditional model is compared to the PBT system which was implemented at the Chair of Computer Science 6, RWTH Aachen University [12]. Moreover, a model combination of the PBT model and the conditional model is evaluated.

7.3 Chinese-English

To obtain the optimal window length of the reordering graph, we performed experiments on the development corpus with different window lengths. Fig. 4 shows the influence of the window length on the translation quality. It points out the importance of reordering for the Chinese-English language pair: a non-monotone translation leads to significantly better results than a monotone translation ($l = 1$). For window lengths higher than $l = 7$, the search space becomes too large to be processed. Heavy pruning had to be applied, which led to a decrease in translation quality.

The results obtained on the Chinese-English test data are summarized in Table 2. A model order of $m = 8$ was used.

The conditional model does not achieve the same translation quality as the PBT model. One reason is the cost function of reordering graphs. The assumption that a monotonic translation is more probable than a translation in which many words are reordered does not hold for the Chinese-English language pair, because the sentence structure between the two languages is often very different. The model combination of the PBT model and the conditional translation model led to an improvement on the test corpus of 0.5 BLEU and 0.7 TER absolute.

7.4 Italian-English

For the Italian-English language pair, the reordering problem is not as pronounced as for the Chinese-English case. Optimal results on the development corpus were obtained for local reorderings with a window length of $l = 3$ and a model order of $m = 4$.

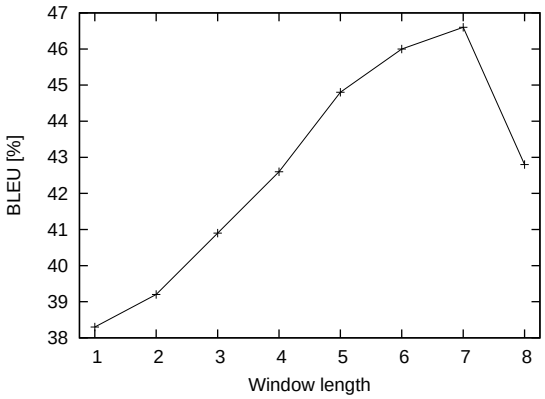


Fig. 4. Translation quality of the development corpus for different reordering window lengths

Table 2. Chinese-English results on the test set

	BLEU [%]	TER [%]
PBT model	39.0	45.1
Monotone translation	21.9	61.1
Conditional model ($l = 7$)	30.6	57.0
Model combination	39.5	44.4

Again, the model combination of the PBT model and the conditional model leads to an improvement in translation quality, though it is not as pronounced as in the case of the Chinese-English corpus.

Table 3. Italian-English results on the test set

	BLEU [%]	TER [%]
PBT model	34.6	50.6
Conditional model ($l = 3$)	31.8	56.5
Model combination	34.9	50.4

8 Conclusion and Outlook

We presented a translation model which takes into account a context beyond phrase boundaries. The conditional model uses a bilingual context to overcome the deficiencies

of the phrase-based translation model. In the Chinese-English case, a complex reordering has to be considered to account for the different sentence structure. Here, reordering phrases is more promising than reordering single words. A model combination which takes advantage of phrase reordering as well as the extension of the context beyond phrase boundaries led to an improvement in performance on both the Chinese-English and the Italian-English corpus.

In the future, advanced smoothing methods should be applied to the conditional model. Moreover, a better reordering model should be developed to take into account differences in word order.

References

1. Brown, P., Della Pietra, S., Della Pietra, V., Mercer, R.: The Mathematics of Statistical Machine Translation: Parameter Estimation Computational Linguistics, Vol 19.2, pages 263-311, 1993.
2. Zens, R., Och F., Ney, H.: Phrase-based statistical machine translation In M. Jarke, J. Koehler, and G. Lakemeyer, editors, 25th German Conf. on Artificial Intelligence (KI2002), volume 2479 of Lecture Notes in Artificial Intelligence (LNAI), pages 18–32, Aachen, Germany, September.
3. Casacuberta, F., Vidal, E.: Machine Translation with Inferred Stochastic Finite-State Transducers, in COLING 2004, Vol. 30, No. 2, pages 205–225, Cambridge, MA, USA
4. Och, F., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). (2002) pp. 295-302, Philadelphia, PA, July.
5. Bakis, R.: Continuous speech word recognition via centisecond acoustic states in Proc. ASA Meeting, Washington DC, 1976, April.
6. Ney, H., Martin, S., Wessel, F.: Statistical Language Modeling Using Leaving-One-Out, Corpus-Based Methods in Language and Speech Processing, pages 174-207, 1997
7. Kanthak, S., Vilar, D., Matusov, E., Zens, R., Ney, H.: Novel Reordering Approaches in Phrase-Based Statistical Machine Translation, in ACL Workshop on Building and Using Parallel Texts, pages 167–174, Association for Computational Linguistics, Ann Arbor, Michigan, June 2005.
8. Snover, M., Dorr, B., Schwartz, R., Makhoul, J., Micciula, L., Weischedel, R.: A Study of Translation Error Rate with Targeted Human Annotation. University of Maryland, College Park and BBN Technologies 2005, July
9. Papineni, K., Roukos, S. ,Ward T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation Technical Report RC22176 (W0109-022) IBM Research Division, Thomas J. Watson Research Center (2001)
10. Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., Yamamoto, S.: Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), pages 147-152
11. Fordyce, C.: Overview of the IWSLT 2007 evaluation campaign International Workshop on Spoken Language Translation (IWSLT) 2007 Trento, Italy, October
12. Mauser, A., Vilar, D., Leusch, G., Zhang, Y., Ney, H.: The RWTH Machine Translation System for IWSLT 2007 International Workshop on Spoken Language Translation, pages 161-168, Trento, Italy, 2007, October.