Discriminative Reordering Models for Statistical Machine Translation

Richard Zens and Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6 – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{zens,ney}@cs.rwth-aachen.de

Abstract

We present discriminative reordering models for phrase-based statistical machine translation. The models are trained using the maximum entropy principle. We use several types of features: based on words, based on word classes, based on the local context. We evaluate the overall performance of the reordering models as well as the contribution of the individual feature types on a word-aligned corpus. Additionally, we show improved translation performance using these reordering models compared to a state-of-the-art baseline system.

1 Introduction

In recent evaluations, phrase-based statistical machine translation systems have achieved good performance. Still the fluency of the machine translation output leaves much to desire. One reason is that most phrase-based systems use a very simple reordering model. Usually, the costs for phrase movements are linear in the distance, e.g. see (Och et al., 1999; Koehn, 2004; Zens et al., 2005).

Recently, in (Tillmann and Zhang, 2005) and in (Koehn et al., 2005), a reordering model has been described that tries to predict the orientation of a phrase, i.e. it answers the question 'should the next phrase be to the left or to the right of the current phrase?' This phrase orientation probability is conditioned on the current source and target phrase and relative frequencies are used to estimate the probabilities.

We adopt the idea of predicting the orientation, but we propose to use a maximum-entropy based model. The relative-frequency based approach may suffer from the data sparseness problem, because most of the phrases occur only once in the training corpus. Our approach circumvents this problem by using a combination of phrase-level and word-level features and by using word-classes or part-of-speech information. Maximum entropy is a suitable framework for combining these different features with a well-defined training criterion.

In (Koehn et al., 2005) several variants of the orientation model have been tried. It turned out that for different tasks, different models show the best performance. Here, we let the maximum entropy training decide which features are important and which features can be neglected. We will see that additional features do not hurt performance and can be safely added to the model.

The remaining part is structured as follows: first we will describe the related work in Section 2 and give a brief description of the baseline system in Section 3. Then, we will present the discriminative reordering model in Section 4. Afterwards, we will evaluate the performance of this new model in Section 5. This evaluation consists of two parts: first we will evaluate the prediction capabilities of the model on a word-aligned corpus and second we will show improved translation quality compared to the baseline system. Finally, we will conclude in Section 6.

2 Related Work

As already mentioned in Section 1, many current phrase-based statistical machine translation systems use a very simple reordering model: the costs

Proceedings of the Workshop on Statistical Machine Translation, pages 55–63,
New York City, June 2006. ©2006 Association for Computational Linguistics

for phrase movements are linear in the distance. This approach is also used in the publicly available Pharaoh decoder (Koehn, 2004). The idea of predicting the orientation is adopted from (Tillmann and Zhang, 2005) and (Koehn et al., 2005). Here, we use the maximum entropy principle to combine a variety of different features.

A reordering model in the framework of weighted finite state transducers is described in (Kumar and Byrne, 2005). There, the movements are defined at the phrase level, but the window for reordering is very limited. The parameters are estimated using an EM-style method.

None of these methods try to generalize from the words or phrases by using word classes or part-of-speech information.

The approach presented here has some resemblance to the bracketing transduction grammars (BTG) of (Wu, 1997), which have been applied to a phrase-based machine translation system in (Zens et al., 2004). The difference is that, here, we do not constrain the phrase reordering. Nevertheless the inverted/monotone concatenation of phrases in the BTG framework is similar to the left/right phrase orientation used here.

3 Baseline System

In statistical machine translation, we are given a source language sentence $f_1^J = f_1 \ldots f_j \ldots f_J$, which is to be translated into a target language sentence $e_1^I = e_1 \ldots e_i \ldots e_I$. Among all possible target language sentences, we will choose the sentence with the highest probability:

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

The posterior probability $Pr(e_1^I | f_1^J)$ is modeled directly using a log-linear combination of several models (Och and Ney, 2002):

$$Pr(e_1^I | f_1^J) = \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum_{I', e'^{I'}_1} \exp\left(\sum_{m=1}^M \lambda_m h_m(e'^{I'}_1, f_1^J)\right)} \quad (2)$$

The denominator represents a normalization factor that depends only on the source sentence f_1^J. Therefore, we can omit it during the search process. As a decision rule, we obtain:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}} \left\{ \sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J) \right\} \quad (3)$$

This approach is a generalization of the source-channel approach (Brown et al., 1990). It has the advantage that additional models $h(\cdot)$ can be easily integrated into the overall system. The model scaling factors λ_1^M are trained with respect to the final translation quality measured by an error criterion (Och, 2003).

We use a state-of-the-art phrase-based translation system (Zens and Ney, 2004; Zens et al., 2005) including the following models: an n-gram language model, a phrase translation model and a word-based lexicon model. The latter two models are used for both directions: $p(f|e)$ and $p(e|f)$. Additionally, we use a word penalty and a phrase penalty. The reordering model of the baseline system is distance-based, i.e. it assigns costs based on the distance from the end position of a phrase to the start position of the next phrase. This very simple reordering model is widely used, for instance in (Och et al., 1999; Koehn, 2004; Zens et al., 2005).

4 The Reordering Model

4.1 Idea

In this section, we will describe the proposed discriminative reordering model.

To make use of word level information, we need the word alignment within the phrase pairs. This can be easily stored during the extraction of the phrase pairs from the bilingual training corpus. If there are multiple possible alignments for a phrase pair, we use the most frequent one.

The notation is introduced using the illustration in Figure 1. There is an example of a left and a right phrase orientation. We assume that we have already produced the three-word phrase in the lower part. Now, the model has to predict if the start position of the next phrase j' is to the left or to the right of the current phrase. The reordering model is applied only at the phrase boundaries. We assume that the reordering within the phrases is correct.

In the remaining part of this section, we will describe the details of this reordering model. The classes our model predicts will be defined in Section 4.2. Then, the feature functions will be defined

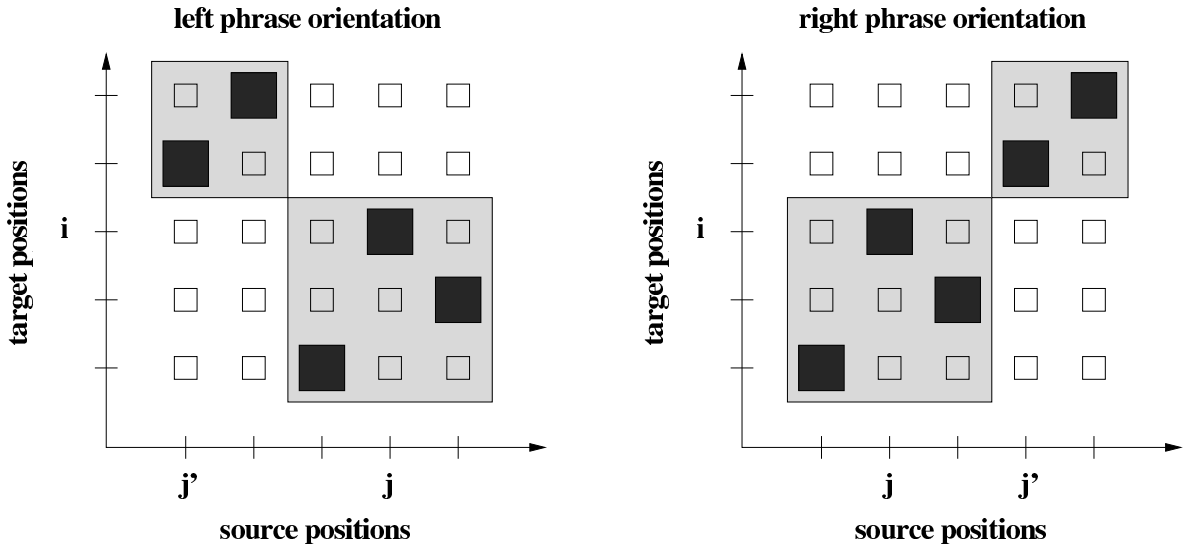

Figure 1: Illustration of the phrase orientation.

in Section 4.3. The training criterion and the training events of the maximum entropy model will be described in Section 4.4.

4.2 Class Definition

Ideally, this model predicts the start position of the next phrase. But as predicting the exact position is rather difficult, we group the possible start positions into classes. In the simplest case, we use only two classes. One class for the positions to the left and one class for the positions to the right. As a refinement, we can use four classes instead of two: 1) one position to the left, 2) more than one positions to the left, 3) one position to the right, 4) more than one positions to the right.

In general, we use a parameter D to specify $2 \cdot D$ classes of the types:

- exactly d positions to the left, $d = 1, ..., D-1$
- at least D positions to the left
- exactly d positions to the right, $d = 1, ..., D-1$
- at least D positions to the right

Let $c_{j,j'}$ denote the orientation class for a movement from source position j to source position j' as illustrated in Figure 1. In the case of two orientation classes, $c_{j,j'}$ is defined as:

$$c_{j,j'} = \begin{cases} \text{left}, & \text{if } j' < j \\ \text{right}, & \text{if } j' > j \end{cases} \quad (4)$$

Then, the reordering model has the form

$$p(c_{j,j'}|f_1^J, e_1^I, i, j)$$

A well-founded framework for directly modeling the probability $p(c_{j,j'}|f_1^J, e_1^I, i, j)$ is maximum entropy (Berger et al., 1996). In this framework, we have a set of N feature functions $h_n(f_1^J, e_1^I, i, j, c_{j,j'}), n = 1, \ldots, N$. Each feature function h_n is weighted with a factor λ_n. The resulting model is:

$$p_{\lambda_1^N}(c_{j,j'}|f_1^J, e_1^I, i, j)$$

$$= \frac{\exp\left(\sum_{n=1}^{N} \lambda_n h_n(f_1^J, e_1^I, i, j, c_{j,j'})\right)}{\sum_{c'} \exp\left(\sum_{n=1}^{N} \lambda_n h_n(f_1^J, e_1^I, i, j, c')\right)} \quad (5)$$

The functional form is identical to Equation 2, but here we will use a large number of binary features, whereas in Equation 2 usually only a very small number of real-valued features is used. More precisely, the resulting reordering model $p_{\lambda_1^N}(c_{j,j'}|f_1^J, e_1^I, i, j)$ is used as an additional component in the log-linear combination of Equation 2.

4.3 Feature Definition

The feature functions of the reordering model depend on the last alignment link (j, i) of a phrase. Note that the source position j is not necessarily the

end position of the source phrase. We use the source position j which is aligned to the last word of the target phrase in target position i. The illustration in Figure 1 contains such an example.

To introduce generalization capabilities, some of the features will depend on word classes or part-of-speech information. Let F_1^J denote the word class sequence that corresponds to the source language sentence f_1^J and let E_1^I denote the target word class sequence that corresponds to the target language sentence e_1^I. Then, the feature functions are of the form $h_n(f_1^J, e_1^I, F_1^J, E_1^I, i, j, j')$. We consider the following binary features:

1. source words within a window around the current source position j

$$h_{f,d,c}(f_1^J, e_1^I, F_1^J, E_1^I, i, j, j') = \delta(f_{j+d}, f) \cdot \delta(c, c_{j,j'}) \quad (6)$$

2. target words within a window around the current target position i

$$h_{e,d,c}(f_1^J, e_1^I, F_1^J, E_1^I, i, j, j') = \delta(e_{i+d}, e) \cdot \delta(c, c_{j,j'}) \quad (7)$$

3. word classes or part-of-speech within a window around the current source position j

$$h_{F,d,c}(f_1^J, e_1^I, F_1^J, E_1^I, i, j, j') = \delta(F_{j+d}, F) \cdot \delta(c, c_{j,j'}) \quad (8)$$

4. word classes or part-of-speech within a window around the current target position i

$$h_{E,d,c}(f_1^J, e_1^I, F_1^J, E_1^I, i, j, j') = \delta(E_{i+d}, E) \cdot \delta(c, c_{j,j'}) \quad (9)$$

Here, $\delta(\cdot,\cdot)$ denotes the Kronecker-function. In the experiments, we will use $d \in \{-1, 0, 1\}$. Many other feature functions are imaginable, e.g. combinations of the described feature functions, n-gram or multi-word features, joint source and target language feature functions.

4.4 Training

As training criterion, we use the maximum class posterior probability. This corresponds to maximizing the likelihood of the maximum entropy model. Since the optimization criterion is convex, there is only a single optimum and no convergence problems occur. To train the model parameters λ_1^N, we use the Generalized Iterative Scaling (GIS) algorithm (Darroch and Ratcliff, 1972).

In practice, the training procedure tends to result in an overfitted model. To avoid overfitting, (Chen and Rosenfeld, 1999) have suggested a smoothing method where a Gaussian prior distribution of the parameters is assumed.

This method tried to avoid very large lambda values and prevents features that occur only once for a specific class from getting a value of infinity.

We train IBM Model 4 with GIZA++ (Och and Ney, 2003) in both translation directions. Then the alignments are symmetrized using a refined heuristic as described in (Och and Ney, 2003). This word-aligned bilingual corpus is used to train the reordering model parameters, i.e. the feature weights λ_1^N. Each alignment link defines an event for the maximum entropy training. An exception are the one-to-many alignments, i.e. one source word is aligned to multiple target words. In this case, only the topmost alignment link is considered because the other ones cannot occur at a phrase boundary. Many-to-one and many-to-many alignments are handled in a similar way.

5 Experimental Results

5.1 Statistics

The experiments were carried out on the *Basic Travel Expression Corpus* (BTEC) task (Takezawa et al., 2002). This is a multilingual speech corpus which contains tourism-related sentences similar to those that are found in phrase books. We use the Arabic-English, the Chinese-English and the Japanese-English data. The corpus statistics are shown in Table 1.

As the BTEC is a rather clean corpus, the preprocessing consisted mainly of tokenization, i.e., separating punctuation marks from words. Additionally, we replaced contractions such as *it's* or *I'm* in the English corpus and we removed the case information. For Arabic, we removed the diacritics and we split common prefixes: Al, w, f, b, l. There was no special preprocessing for the Chinese and the Japanese training corpora.

To train and evaluate the reordering model, we

Table 1: Corpus statistics after preprocessing for the BTEC task.

		Arabic	Chinese	Japanese	English
Train	Sentences	20 000			
	Running Words	180 075	176 199	198 453	189 927
	Vocabulary	15 371	8 687	9 277	6 870
C-Star'03	Sentences	506			
	Running Words	3 552	3 630	4 130	3 823

Table 2: Statistics of the training and test word alignment links.

	Ara-Eng	Chi-Eng	Jap-Eng
Training	144K	140K	119K
Test	16.2K	15.7K	13.2K

use the word aligned bilingual training corpus. For evaluating the classification power of the reordering model, we partition the corpus into a training part and a test part. In our experiments, we use about 10% of the corpus for testing and the remaining part for training the feature weights of the reordering model with the GIS algorithm using YASMET (Och, 2001). The statistics of the training and test alignment links is shown in Table 2. The number of training events ranges from 119K for Japanese-English to 144K for Arabic-English.

The word classes for the class-based features are trained using the `mkcls` tool (Och, 1999). In the experiments, we use 50 word classes. Alternatively, one could use part-of-speech information for this purpose.

Additional experiments were carried out on the large data track of the Chinese-English NIST task. The corpus statistics of the bilingual training corpus are shown in Table 3. The language model was trained on the English part of the bilingual training corpus and additional monolingual English data from the GigaWord corpus. The total amount of language model training data was about 600M running words. We use a fourgram language model with modified Kneser-Ney smoothing as implemented in the SRILM toolkit (Stolcke, 2002). For the four English reference translations of the evaluation sets, the accumulated statistics are presented.

Table 3: Chinese-English NIST task: corpus statistics for the bilingual training data and the NIST evaluation sets of the years 2002 to 2005.

			Chinese	English
Train	Sentence Pairs		7M	
	Running Words		199M	213M
	Vocabulary Size		223K	351K
	Dictionary Entry Pairs		82K	
Eval	2002	Sentences	878	3 512
		Running Words	25K	105K
	2003	Sentences	919	3 676
		Running Words	26K	122K
	2004	Sentences	1788	7 152
		Running Words	52K	245K
	2005	Sentences	1082	4 328
		Running Words	33K	148K

5.2 Classification Results

In this section, we present the classification results for the three language pairs. In Table 4, we present the classification results for two orientation classes.

As baseline we always choose the most frequent orientation class. For Arabic-English, the baseline is with 6.3% already very low. This means that the word order in Arabic is very similar to the word order in English. For Chinese-English, the baseline is with 12.7% about twice as large. The most differences in word order occur for Japanese-English. This seems to be reasonable as Japanese has usually a different sentence structure, subject-object-verb compared to subject-verb-object in English.

For each language pair, we present results for several combination of features. The three columns per language pair indicate if the features are based on the words (column label 'Words'), on the word classes (column label 'Classes') or on both (column label

Table 4: Classification error rates [%] using two orientation classes.

		Arabic-English			Chinese-English			Japanese-English		
Baseline		6.3			12.7			26.2		
Lang.	Window	Words	Classes	W+C	Words	Classes	W+C	Words	Classes	W+C
Tgt	$d = 0$	4.7	5.3	4.4	9.3	10.4	8.9	13.6	15.1	13.4
	$d \in \{0, 1\}$	4.5	5.0	4.3	8.9	9.9	8.6	13.7	14.9	13.4
	$d \in \{-1, 0, 1\}$	4.5	4.9	4.3	8.6	9.5	8.3	13.5	14.6	13.3
Src	$d = 0$	5.6	5.0	3.9	7.9	8.3	7.2	12.2	11.8	11.0
	$d \in \{0, 1\}$	3.2	3.0	2.6	4.7	4.7	4.2	10.1	9.7	9.4
	$d \in \{-1, 0, 1\}$	2.9	2.5	2.3	3.9	3.5	3.3	9.0	8.0	7.8
Src	$d = 0$	4.3	3.9	3.7	7.1	7.8	6.5	10.8	10.9	9.8
+	$d \in \{0, 1\}$	2.9	2.6	2.5	4.6	4.5	4.1	9.3	9.1	8.6
Tgt	$d \in \{-1, 0, 1\}$	2.8	2.1	2.1	3.9	3.4	3.3	8.7	7.7	7.7

'W+C'). We also distinguish if the features depend on the target sentence ('Tgt'), on the source sentence ('Src') or on both ('Src+Tgt').

For Arabic-English, using features based only on words of the target sentence the classification error rate can be reduced to 4.5%. If the features are based only on the source sentence words, a classification error rate of 2.9% is reached. Combining the features based on source and target sentence words, a classification error rate of 2.8% can be achieved. Adding the features based on word classes, the classification error rate can be further improved to 2.1%. For the other language pairs, the results are similar except that the absolute values of the classification error rates are higher.

We observe the following:

- The features based on the source sentence perform better than features based on the target sentence.
- Combining source and target sentence features performs best.
- Increasing the window always helps, i.e. additional context information is useful.
- Often the word-class based features outperform the word-based features.
- Combining word-based and word-class based features performs best.
- In general, adding features does not hurt the performance.

These are desirable properties of an appropriate reordering model. The main point is that these are fulfilled not only on the training data, but on unseen test data. There seems to be no overfitting problem.

In Table 5, we present the results for four orientation classes. The final error rates are a factor 2-4 larger than for two orientation classes. Despite that we observe the same tendencies as for two orientation classes. Again, using more features always helps to improve the performance.

5.3 Translation Results

For the translation experiments on the BTEC task, we report the two accuracy measures BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) as well as the two error rates: word error rate (WER) and position-independent word error rate (PER). These criteria are computed with respect to 16 references.

In Table 6, we show the translation results for the BTEC task. In these experiments, the reordering model uses two orientation classes, i.e. it predicts either a left or a right orientation. The features for the maximum-entropy based reordering model are based on the source and target language words within a window of one. The word-class based features are not used for the translation experiments. The maximum-entropy based reordering model achieves small but consistent improvement for all the evaluation criteria. Note that the baseline system, i.e. using the distance-based reordering, was among the best systems in the IWSLT 2005 evalua-

Table 5: Classification error rates [%] using four orientation classes.

		Arabic-English			Chinese-English			Japanese-English		
Baseline		31.4			44.9			59.0		
Lang.	Window	Words	Classes	W+C	Words	Classes	W+C	Words	Classes	W+C
Tgt	$d = 0$	24.5	27.7	24.2	30.0	34.4	29.7	28.9	31.4	28.7
	$d \in \{0, 1\}$	23.9	27.2	23.7	29.2	32.9	28.9	28.7	30.6	28.3
	$d \in \{-1, 0, 1\}$	22.1	25.3	21.9	27.6	31.4	27.4	28.3	30.1	28.2
Src	$d = 0$	22.1	23.2	20.4	25.9	27.7	20.4	24.1	24.9	22.3
	$d \in \{0, 1\}$	11.9	12.0	10.8	14.0	14.9	13.2	18.6	19.5	17.7
	$d \in \{-1, 0, 1\}$	10.1	8.7	8.0	11.4	11.1	10.5	15.6	15.6	14.5
Src	$d = 0$	20.9	21.8	19.6	24.1	26.8	19.6	22.3	23.4	21.1
+	$d \in \{0, 1\}$	11.8	11.5	10.6	13.5	14.5	12.8	18.6	18.8	17.1
Tgt	$d \in \{-1, 0, 1\}$	9.6	7.7	7.6	11.3	10.1	10.1	15.6	15.2	14.2

Table 6: Translation Results for the BTEC task.

Language Pair	Reordering	WER [%]	PER [%]	NIST	BLEU [%]
Arabic-English	Distance-based	24.1	20.9	10.0	63.8
	Max-Ent based	23.6	20.7	10.1	64.8
Chinese-English	Distance-based	50.4	43.0	7.67	44.4
	Max-Ent based	49.3	42.4	7.36	45.8
Japanese-English	Distance-based	32.1	25.2	8.96	56.2
	Max-Ent based	31.2	25.2	9.00	56.8

tion campaign (Eck and Hori, 2005).

Some translation examples are presented in Table 7. We observe that the system using the maximum-entropy based reordering model produces more fluent translations.

Additional translation experiments were carried out on the large data track of the Chinese-English NIST task. For this task, we use only the BLEU and NIST scores. Both scores are computed case-insensitive with respect to four reference translations using the mteval-v11b tool[1].

For the NIST task, we use the BLEU score as primary criterion which is optimized on the NIST 2002 evaluation set using the Downhill Simplex algorithm (Press et al., 2002). Note that only the eight or nine model scaling factors of Equation 2 are optimized using the Downhill Simplex algorithm. The feature weights of the reordering model are trained using the GIS algorithm as described in Section 4.4. We use a state-of-the-art baseline system which would have obtained a good rank in the last NIST evaluation (NIST, 2005).

The translation results for the NIST task are presented in Table 8. We observe consistent improvements of the BLEU score on all evaluation sets. The overall improvement due to reordering ranges from 1.2% to 2.0% absolute. The contribution of the maximum-entropy based reordering model to this improvement is in the range of 25% to 58%, e.g. for the NIST 2003 evaluation set about 58% of the improvement using reordering can be attributed to the maximum-entropy based reordering model.

We also measured the classification performance for the NIST task. The general tendencies are identical to the BTEC task.

6 Conclusions

We have presented a novel discriminative reordering model for statistical machine translation. This model is trained on the word aligned bilingual corpus using the maximum entropy principle. Several types of features have been used:

- based on the source and target sentence

[1] http://www.nist.gov/speech/tests/mt/resources/scoring.htm

Table 7: Translation examples for the BTEC task.

System	Translation
Distance-based	I would like to check out time one day before.
Max-Ent based	I would like to check out one day before the time.
Reference	I would like to check out one day earlier.
Distance-based	I hate pepper green.
Max-Ent based	I hate the green pepper.
Reference	I hate green peppers.
Distance-based	Is there a subway map where?
Max-Ent based	Where is the subway route map?
Reference	Where do they have a subway map?

Table 8: Translation results for several evaluation sets of the Chinese-English NIST task.

Evaluation set	2002 (dev)		2003		2004		2005	
Reordering	NIST	BLEU[%]	NIST	BLEU[%]	NIST	BLEU[%]	NIST	BLEU[%]
None	8.96	33.5	8.67	32.7	8.76	32.0	8.62	30.8
Distance-based	9.19	34.6	8.85	33.2	9.05	33.2	8.79	31.6
Max-Ent based	9.24	35.5	8.87	33.9	9.04	33.6	8.78	32.1

- based on words and word classes
- using local context information

We have evaluated the performance of the reordering model on a held-out word-aligned corpus. We have shown that the model is able to predict the orientation very well, e.g. for Arabic-English the classification error rate is only 2.1%.

We presented improved translation results for three language pairs on the BTEC task and for the large data track of the Chinese-English NIST task.

In none of the cases additional features have hurt the classification performance on the held-out test corpus. This is a strong evidence that the maximum entropy framework is suitable for this task.

Another advantage of our approach is the generalization capability via the use of word classes or part-of-speech information. Furthermore, additional features can be easily integrated into the maximum entropy framework.

So far, the word classes were not used for the translation experiments. As the word classes help for the classification task, we might expect further improvements of the translation results. Using part-of-speech information instead (or in addition) to the automatically computed word classes might also be beneficial. More fine-tuning of the reordering model toward translation quality might also result in improvements. As already mentioned in Section 4.3, a richer feature set could be helpful.

Acknowledgments

This material is partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023, and was partly funded by the European Union under the integrated project TC-STAR (Technology and Corpora for Speech to Speech Translation, IST-2002-FP6-506738, http://www.tc-star.org).

References

A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72, March.

P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June.

S. F. Chen and R. Rosenfeld. 1999. A gaussian prior for smoothing maximum entropy models. Technical Report CMUCS-99-108, Carnegie Mellon University, Pittsburgh, PA.

J. N. Darroch and D. Ratcliff. 1972. Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43:1470–1480.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. ARPA Workshop on Human Language Technology*.

M. Eck and C. Hori. 2005. Overview of the IWSLT 2005 evaluation campaign. In *Proc. International Workshop on Spoken Language Translation (IWSLT)*, Pittsburgh, PA, October.

P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proc. International Workshop on Spoken Language Translation (IWSLT)*, Pittsburgh, PA, October.

P. Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proc. 6th Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, pages 115–124, Washington DC, September/October.

S. Kumar and W. Byrne. 2005. Local phrase reordering models for statistical machine translation. In *Proc. of the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 161–168, Vancouver, Canada, October.

NIST. 2005. NIST 2005 machine translation evaluation official results. `http://www.nist.gov/speech/tests/mt/mt05eval_official_results_release_20050801_v3.html`, August.

F. J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, July.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

F. J. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June.

F. J. Och. 1999. An efficient method for determining bilingual word classes. In *Proc. 9th Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL)*, pages 71–76, Bergen, Norway, June.

F. J. Och. 2001. YASMET: Toolkit for conditional maximum entropy models. http://www-i6.informatik.rwth-aachen.de/web/Software/YASMET.html.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 2002. *Numerical Recipes in C++*. Cambridge University Press, Cambridge, UK.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO.

T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of the Third Int. Conf. on Language Resources and Evaluation (LREC)*, pages 147–152, Las Palmas, Spain, May.

C. Tillmann and T. Zhang. 2005. A localized prediction model for statistical machine translation. In *Proc. of the 43rd Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 557–564, Ann Arbor, MI, June.

D. Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403, September.

R. Zens and H. Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proc. Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL)*, pages 257–264, Boston, MA, May.

R. Zens, H. Ney, T. Watanabe, and E. Sumita. 2004. Reordering constraints for phrase-based statistical machine translation. In *Proc. 20th Int. Conf. on Computational Linguistics (COLING)*, pages 205–211, Geneva, Switzerland, August.

R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, and H. Ney. 2005. The RWTH phrase-based statistical machine translation system. In *Proc. International Workshop on Spoken Language Translation (IWSLT)*, pages 155–162, Pittsburgh, PA, October.