Discriminative Training and Maximum Entropy Models for Statistical Machine Translation

Franz Josef Och and **Hermann Ney**
Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen - University of Technology
D-52056 Aachen, Germany
{och,ney}@informatik.rwth-aachen.de

Abstract

We present a framework for statistical machine translation of natural languages based on direct maximum entropy models, which contains the widely used source-channel approach as a special case. All knowledge sources are treated as feature functions, which depend on the source language sentence, the target language sentence and possible hidden variables. This approach allows a baseline machine translation system to be extended easily by adding new feature functions. We show that a baseline statistical machine translation system is significantly improved using this approach.

1 Introduction

We are given a source ('French') sentence $f_1^J = f_1, \ldots, f_j, \ldots, f_J$, which is to be translated into a target ('English') sentence $e_1^I = e_1, \ldots, e_i, \ldots, e_I$. Among all possible target sentences, we will choose the sentence with the highest probability:[1]

$$\hat{e}_1^I = \operatorname*{argmax}_{e_1^I} \{Pr(e_1^I|f_1^J)\} \qquad (1)$$

The argmax operation denotes the search problem, i.e. the generation of the output sentence in the target language.

1.1 Source-Channel Model

According to Bayes' decision rule, we can equivalently to Eq. 1 perform the following maximization:

$$\hat{e}_1^I = \operatorname*{argmax}_{e_1^I} \{Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)\} \qquad (2)$$

This approach is referred to as source-channel approach to statistical MT. Sometimes, it is also referred to as the 'fundamental equation of statistical MT' (Brown et al., 1993). Here, $Pr(e_1^I)$ is the language model of the target language, whereas $Pr(f_1^J|e_1^I)$ is the translation model. Typically, Eq. 2 is favored over the direct translation model of Eq. 1 with the argument that it yields a modular approach. Instead of modeling one probability distribution, we obtain two different knowledge sources that are trained independently.

The overall architecture of the source-channel approach is summarized in Figure 1. In general, as shown in this figure, there may be additional transformations to make the translation task simpler for the algorithm. Typically, training is performed by applying a maximum likelihood approach. If the language model $Pr(e_1^I) = p_\gamma(e_1^I)$ depends on parameters γ and the translation model $Pr(f_1^J|e_1^I) = p_\theta(f_1^J|e_1^I)$ depends on parameters θ, then the optimal parameter values are obtained by maximizing the likelihood on a parallel training corpus $\mathbf{f}_1^S, \mathbf{e}_1^S$ (Brown et al., 1993):

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \prod_{s=1}^{S} p_\theta(\mathbf{f}_s|\mathbf{e}_s) \qquad (3)$$

$$\hat{\gamma} = \operatorname*{argmax}_{\gamma} \prod_{s=1}^{S} p_\gamma(\mathbf{e}_s) \qquad (4)$$

[1] The notational convention will be as follows. We use the symbol $Pr(\cdot)$ to denote general probability distributions with (nearly) no specific assumptions. In contrast, for model-based probability distributions, we use the generic symbol $p(\cdot)$.

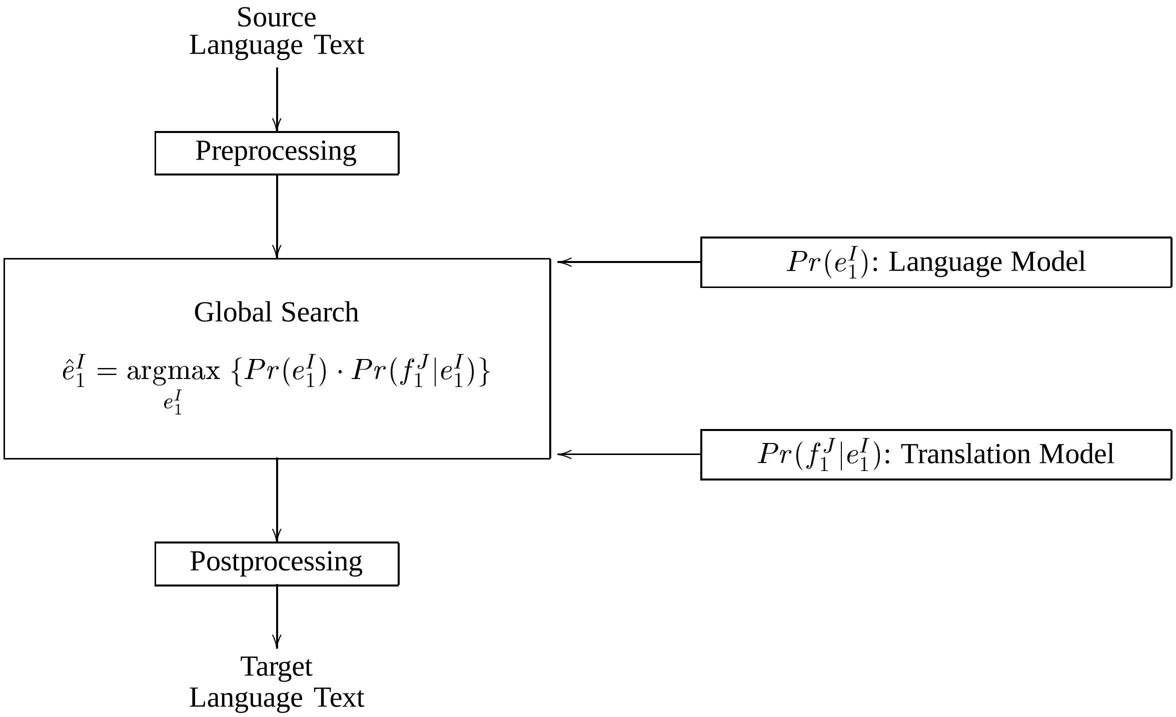
Source Language Text
Preprocessing
Global Search
$\hat{e}_1^I = \underset{e_1^I}{\text{argmax}} \; \{Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)\}$
$Pr(e_1^I)$: Language Model
$Pr(f_1^J|e_1^I)$: Translation Model
Postprocessing
Target Language Text

Figure 1: Architecture of the translation approach based on source-channel models.

We obtain the following decision rule:

$$\hat{e}_1^I = \underset{e_1^I}{\text{argmax}}\{p_{\hat{\gamma}}(e_1^I) \cdot p_{\hat{\theta}}(f_1^J|e_1^I)\} \qquad (5)$$

State-of-the-art statistical MT systems are based on this approach. Yet, the use of this decision rule has various problems:

1. The combination of the language model $p_{\hat{\gamma}}(e_1^I)$ and the translation model $p_{\hat{\theta}}(f_1^J|e_1^I)$ as shown in Eq. 5 can only be shown to be optimal if the true probability distributions $p_{\hat{\gamma}}(e_1^I) = Pr(e_1^I)$ and $p_{\hat{\theta}}(f_1^J|e_1^I) = Pr(f_1^J|e_1^I)$ are used. Yet, we know that the used models and training methods provide only poor approximations of the true probability distributions. Therefore, a different combination of language model and translation model might yield better results.

2. There is no straightforward way to extend a baseline statistical MT model by including additional dependencies.

3. Often, we observe that comparable results are obtained by using the following decision rule instead of Eq. 5 (Och et al., 1999):

$$\hat{e}_1^I = \underset{e_1^I}{\text{argmax}}\{p_{\hat{\gamma}}(e_1^I) \cdot p_{\hat{\theta}}(e_1^I|f_1^J)\} \qquad (6)$$

Here, we replaced $p_{\hat{\theta}}(f_1^J|e_1^I)$ by $p_{\hat{\theta}}(e_1^I|f_1^J)$. From a theoretical framework of the source-channel approach, this approach is hard to justify. Yet, if both decision rules yield the same translation quality, we can use that decision rule which is better suited for efficient search.

1.2 Direct Maximum Entropy Translation Model

As alternative to the source-channel approach, we directly model the posterior probability $Pr(e_1^I|f_1^J)$. An especially well-founded framework for doing this is maximum entropy (Berger et al., 1996). In this framework, we have a set of M feature functions $h_m(e_1^I, f_1^J), m = 1, \ldots, M$. For each feature function, there exists a model parameter $\lambda_m, m = 1, \ldots, M$. The direct translation probability is given

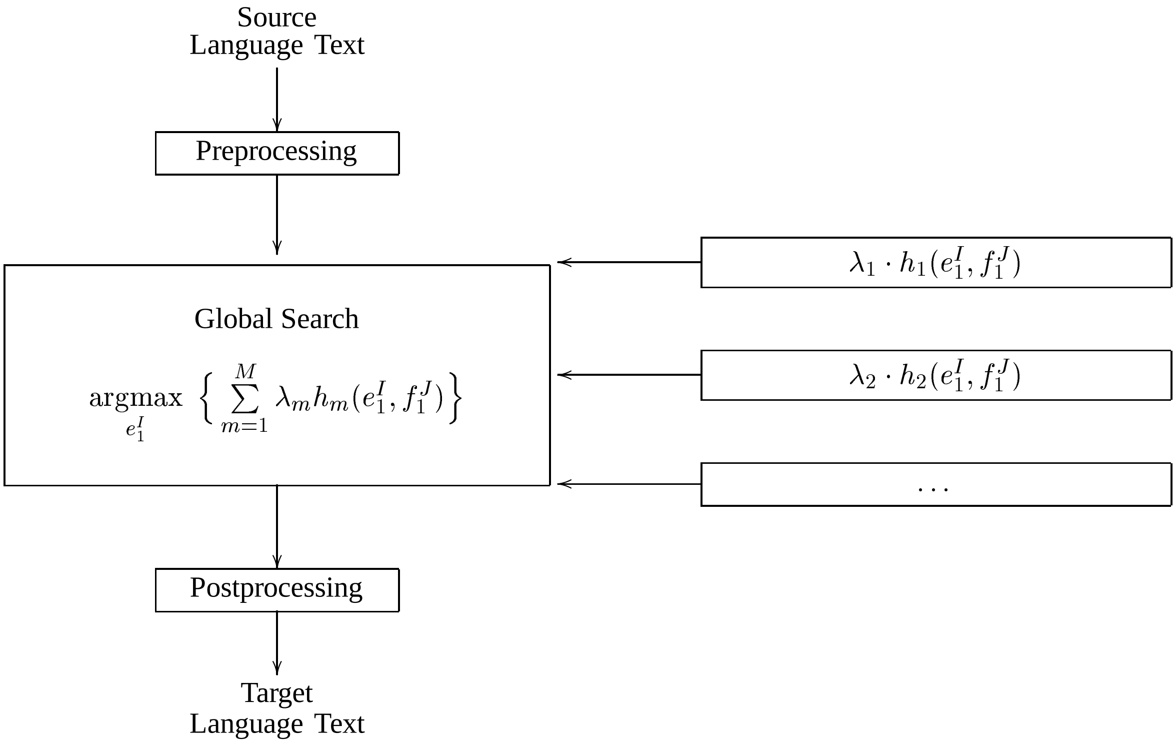
Source Language Text
Preprocessing
Global Search
$\underset{e_1^I}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\}$
$\lambda_1 \cdot h_1(e_1^I, f_1^J)$
$\lambda_2 \cdot h_2(e_1^I, f_1^J)$
$\ldots$
Postprocessing
Target Language Text

Figure 2: Architecture of the translation approach based on direct maximum entropy models.

by:

$$\begin{aligned} Pr(e_1^I|f_1^J) &= p_{\lambda_1^M}(e_1^I|f_1^J) && (7) \\ &= \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)]}{\sum_{e'^I_1} \exp[\sum_{m=1}^{M} \lambda_m h_m(e'^I_1, f_1^J)]} && (8) \end{aligned}$$

This approach has been suggested by (Papineni et al., 1997; Papineni et al., 1998) for a natural language understanding task.

We obtain the following decision rule:

$$\begin{aligned} \hat{e}_1^I &= \underset{e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I|f_1^J) \right\} \\ &= \underset{e_1^I}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\} \end{aligned}$$

Hence, the time-consuming renormalization in Eq. 8 is not needed in search. The overall architecture of the direct maximum entropy models is summarized in Figure 2.

Interestingly, this framework contains as special case the source channel approach (Eq. 5) if we use the following two feature functions:

$$\begin{aligned} h_1(e_1^I, f_1^J) &= \log p_{\hat{\gamma}}(e_1^I) && (9) \\ h_2(e_1^I, f_1^J) &= \log p_{\hat{\theta}}(f_1^J|e_1^I) && (10) \end{aligned}$$

and set $\lambda_1 = \lambda_2 = 1$. Optimizing the corresponding parameters λ_1 and λ_2 of the model in Eq. 8 is equivalent to the optimization of model scaling factors, which is a standard approach in other areas such as speech recognition or pattern recognition.

The use of an ‘inverted’ translation model in the unconventional decision rule of Eq. 6 results if we use the feature function $\log Pr(e_1^I|f_1^J)$ instead of $\log Pr(f_1^J|e_1^I)$. In this framework, this feature can be as good as $\log Pr(f_1^J|e_1^I)$. It has to be empirically verified, which of the two features yields better results. We even can use both features $\log Pr(e_1^I|f_1^J)$ and $\log Pr(f_1^J|e_1^I)$, obtaining a more symmetric translation model.

As training criterion, we use the maximum class posterior probability criterion:

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmax}} \left\{ \sum_{s=1}^{S} \log p_{\lambda_1^M}(\mathbf{e}_s|\mathbf{f}_s) \right\} \quad (11)$$

This corresponds to maximizing the equivocation or maximizing the likelihood of the direct translation model. This direct optimization of the posterior probability in Bayes decision rule is referred to as discriminative training (Ney, 1995) because we directly take into account the overlap in the probability distributions. The optimization problem has one global optimum and the optimization criterion is convex.

1.3 Alignment Models and Maximum Approximation

Typically, the probability $Pr(f_1^J|e_1^I)$ is decomposed via additional hidden variables. In statistical alignment models $Pr(f_1^J, a_1^J|e_1^I)$, the alignment a_1^J is introduced as a hidden variable:

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

The alignment mapping is $j \rightarrow i = a_j$ from source position j to target position $i = a_j$.

Search is performed using the so-called maximum approximation:

$$\begin{aligned} \hat{e}_1^I &= \operatorname*{argmax}_{e_1^I} \left\{ Pr(e_1^I) \cdot \sum_{a_1^J} Pr(f_1^J, a_1^J|e_1^I) \right\} \\ &\approx \operatorname*{argmax}_{e_1^I} \left\{ Pr(e_1^I) \cdot \max_{a_1^J} Pr(f_1^J, a_1^J|e_1^I) \right\} \end{aligned}$$

Hence, the search space consists of the set of all possible target language sentences e_1^I and all possible alignments a_1^J.

Generalizing this approach to direct translation models, we extend the feature functions to include the dependence on the additional hidden variable. Using M feature functions of the form $h_m(e_1^I, f_1^J, a_1^J), m = 1, \ldots, M$, we obtain the following model:

$$\begin{aligned} Pr(e_1^I, a_1^J|f_1^J) &= \\ &= \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J, a_1^J)\right)}{\sum_{e'^I_1, a'^J_1} \exp\left(\sum_{m=1}^M \lambda_m h_m(e'^I_1, f_1^J, a'^J_1)\right)} \end{aligned}$$

Obviously, we can perform the same step for translation models with an even richer structure of hidden variables than only the alignment a_1^J. To simplify the notation, we shall omit in the following the dependence on the hidden variables of the model.

2 Alignment Templates

As specific MT method, we use the alignment template approach (Och et al., 1999). The key elements of this approach are the *alignment templates*, which are pairs of source and target language phrases together with an alignment between the words within the phrases. The advantage of the alignment template approach compared to single word-based statistical translation models is that word context and local changes in word order are explicitly considered.

The alignment template model refines the translation probability $Pr(f_1^J|e_1^I)$ by introducing two hidden variables z_1^K and a_1^K for the K alignment templates and the alignment of the alignment templates:

$$\begin{aligned} Pr(f_1^J|e_1^I) &= \sum_{z_1^K, a_1^K} Pr(a_1^K|e_1^I) \cdot \\ & \qquad Pr(z_1^K|a_1^K, e_1^I) \cdot Pr(f_1^J|z_1^K, a_1^K, e_1^I) \end{aligned}$$

Hence, we obtain three different probability distributions: $Pr(a_1^K|e_1^I)$, $Pr(z_1^K|a_1^K, e_1^I)$ and $Pr(f_1^J|z_1^K, a_1^K, e_1^I)$. Here, we omit a detailed description of modeling, training and search, as this is not relevant for the subsequent exposition. For further details, see (Och et al., 1999).

To use these three component models in a direct maximum entropy approach, we define three different feature functions for each component of the translation model instead of one feature function for the whole translation model $p(f_1^J|e_1^I)$. The feature functions have then not only a dependence on f_1^J and e_1^I but also on z_1^K, a_1^K.

3 Feature functions

So far, we use the logarithm of the components of a translation model as feature functions. This is a very convenient approach to improve the quality of a baseline system. Yet, we are not limited to train only model scaling factors, but we have many possibilities:

- We could add a sentence length feature:

  $$h(f_1^J, e_1^I) = I$$

  This corresponds to a word penalty for each produced target word.

- We could use additional language models by using features of the following form:

$$h(f_1^J, e_1^I) = h(e_1^I)$$

- We could use a feature that counts how many entries of a conventional lexicon co-occur in the given sentence pair. Therefore, the weight for the provided conventional dictionary can be learned. The intuition is that the conventional dictionary is expected to be more reliable than the automatically trained lexicon and therefore should get a larger weight.

- We could use lexical features, which fire if a certain lexical relationship (f, e) occurs:

$$h(f_1^J, e_1^I) = \left(\sum_{j=1}^{J} \delta(f, f_j) \right) \cdot \left(\sum_{i=1}^{I} \delta(e, e_i) \right)$$

- We could use grammatical features that relate certain grammatical dependencies of source and target language. For example, using a function $k(\cdot)$ that counts how many verb groups exist in the source or the target sentence, we can define the following feature, which is 1 if each of the two sentences contains the same number of verb groups:

$$h(f_1^J, e_1^I) = \delta(k(f_1^J), k(e_1^I)) \qquad (12)$$

  In the same way, we can introduce semantic features or pragmatic features such as the dialogue act classification.

We can use numerous additional features that deal with specific problems of the baseline statistical MT system. In this paper, we shall use the first three of these features. As additional language model, we use a class-based five-gram language model. This feature and the word penalty feature allow a straightforward integration into the used dynamic programming search algorithm (Och et al., 1999). As this is not possible for the conventional dictionary feature, we use n-best rescoring for this feature.

4 Training

To train the model parameters λ_1^M of the direct translation model according to Eq. 11, we use the GIS (Generalized Iterative Scaling) algorithm (Darroch and Ratcliff, 1972). It should be noted that, as was already shown by (Darroch and Ratcliff, 1972), by applying suitable transformations, the GIS algorithm is able to handle any type of real-valued features. To apply this algorithm, we have to solve various practical problems.

The renormalization needed in Eq. 8 requires a sum over a large number of possible sentences, for which we do not know an efficient algorithm. Hence, we approximate this sum by sampling the space of all possible sentences by a large set of highly probable sentences. The set of considered sentences is computed by an appropriately extended version of the used search algorithm (Och et al., 1999) computing an approximate n-best list of translations.

Unlike automatic speech recognition, we do not have one reference sentence, but there exists a number of reference sentences. Yet, the criterion as it is described in Eq. 11 allows for only one reference translation. Hence, we change the criterion to allow R_s reference translations $\mathbf{e}_{s,1}, \ldots, \mathbf{e}_{s,R_s}$ for the sentence $\mathbf{e}_s$:

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmax}} \left\{ \sum_{s=1}^{S} \frac{1}{R_s} \sum_{r=1}^{R_s} \log p_{\lambda_1^M}(\mathbf{e}_{s,r}|\mathbf{f}_s) \right\}$$

We use this optimization criterion instead of the optimization criterion shown in Eq. 11.

In addition, we might have the problem that no single of the reference translations is part of the n-best list because the search algorithm performs pruning, which in principle limits the possible translations that can be produced given a certain input sentence. To solve this problem, we define for maximum entropy training each sentence as reference translation that has the minimal number of word errors with respect to any of the reference translations.

5 Results

We present results on the VERBMOBIL task, which is a speech translation task in the domain of appointment scheduling, travel planning, and hotel reservation (Wahlster, 1993). Table 1 shows the corpus statistics of this task. We use a training corpus, which is used to train the alignment template model and the language models, a development corpus, which is used to estimate the model scaling factors, and a test corpus.

Table 1: Characteristics of training corpus (Train), manual lexicon (Lex), development corpus (Dev), test corpus (Test).

		German	English
Train	Sentences	58 073	
	Words	519 523	549 921
	Singletons	3 453	1 698
	Vocabulary	7 939	4 672
Lex	Entries	12 779	
	Ext. Vocab.	11 501	6 867
Dev	Sentences	276	
	Words	3 159	3 438
	PP (trigr. LM)	-	28.1
Test	Sentences	251	
	Words	2 628	2 871
	PP (trigr. LM)	-	30.5

So far, in machine translation research does not exist one generally accepted criterion for the evaluation of the experimental results. Therefore, we use a large variety of different criteria and show that the obtained results improve on most or all of these criteria. In all experiments, we use the following six error criteria:

- SER (sentence error rate): The SER is computed as the number of times that the generated sentence corresponds exactly to one of the reference translations used for the maximum entropy training.

- WER (word error rate): The WER is computed as the minimum number of substitution, insertion and deletion operations that have to be performed to convert the generated sentence into the target sentence.

- PER (position-independent WER): A shortcoming of the WER is the fact that it requires a perfect word order. The word order of an acceptable sentence can be different from that of the target sentence, so that the WER measure alone could be misleading. To overcome this problem, we introduce as additional measure the position-independent word error rate (PER). This measure compares the words in the two sentences ignoring the word order.

- mWER (multi-reference word error rate): For each test sentence, there is not only used a single reference translation, as for the WER, but a whole set of reference translations. For each translation hypothesis, the edit distance to the most similar sentence is calculated (Nießen et al., 2000).

- BLEU score: This score measures the precision of unigrams, bigrams, trigrams and fourgrams with respect to a whole set of reference translations with a penalty for too short sentences (Papineni et al., 2001). Unlike all other evaluation criteria used here, BLEU measures accuracy, i.e. the opposite of error rate. Hence, large BLEU scores are better.

- SSER (subjective sentence error rate): For a more detailed analysis, subjective judgments by test persons are necessary. Each translated sentence was judged by a human examiner according to an error scale from 0.0 to 1.0 (Nießen et al., 2000).

- IER (information item error rate): The test sentences are segmented into information items. For each of them, if the intended information is conveyed and there are no syntactic errors, the sentence is counted as correct (Nießen et al., 2000).

In the following, we present the results of this approach. Table 2 shows the results if we use a direct translation model (Eq. 6).

As baseline features, we use a normal word trigram language model and the three component models of the alignment templates. The first row shows the results using only the four baseline features with $\lambda_1 = \cdots = \lambda_4 = 1$. The second row shows the result if we train the model scaling factors. We see a systematic improvement on all error rates. The following three rows show the results if we add the word penalty, an additional class-based five-gram

Table 2: Effect of maximum entropy training for alignment template approach (WP: word penalty feature, CLM: class-based language model (five-gram), MX: conventional dictionary).

	objective criteria [%]					subjective criteria [%]	
	SER	WER	PER	mWER	BLEU	SSER	IER
Baseline($\lambda_m = 1$)	86.9	42.8	33.0	37.7	43.9	35.9	39.0
ME	81.7	40.2	28.7	34.6	49.7	32.5	34.8
ME+WP	80.5	38.6	26.9	32.4	54.1	29.9	32.2
ME+WP+CLM	78.1	38.3	26.9	32.1	55.0	29.1	30.9
ME+WP+CLM+MX	77.8	38.4	26.8	31.9	55.2	28.8	30.9

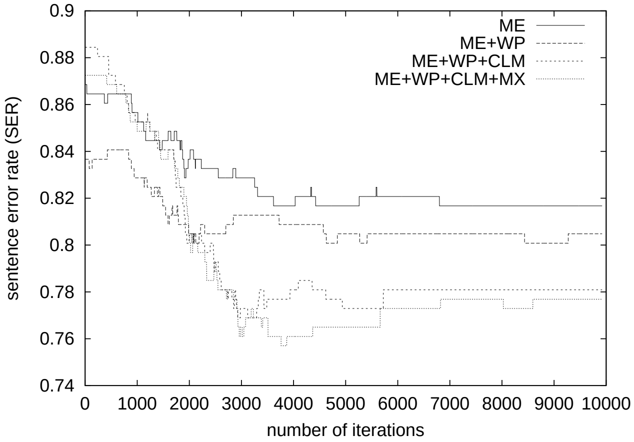

Figure 3: Test error rate over the iterations of the GIS algorithm for maximum entropy training of alignment templates.

Table 3: Resulting model scaling factors of maximum entropy training for alignment templates; λ_1: trigram language model; λ_2: alignment template model, λ_3: lexicon model, λ_4: alignment model (normalized such that $\sum_{m=1}^{4} \lambda_m = 4$).

	ME	+WP	+CLM	+MX
λ_1	0.86	0.98	0.75	0.77
λ_2	2.33	2.05	2.24	2.24
λ_3	0.58	0.72	0.79	0.75
λ_4	0.22	0.25	0.23	0.24
WP	·	2.6	3.03	2.78
CLM	·	·	0.33	0.34
MX	·	·	·	2.92

language model and the conventional dictionary features. We observe improved error rates for using the word penalty and the class-based language model as additional features.

Figure 3 show how the sentence error rate (SER) on the test corpus improves during the iterations of the GIS algorithm. We see that the sentence error rates converges after about 4000 iterations. We do not observe significant overfitting.

Table 3 shows the resulting normalized model scaling factors. Multiplying each model scaling factor by a constant positive value does not affect the decision rule. We see that adding new features also has an effect on the other model scaling factors.

6 Related Work

The use of direct maximum entropy translation models for statistical machine translation has been suggested by (Papineni et al., 1997; Papineni et al., 1998). They train models for natural language understanding rather than natural language translation. In contrast to their approach, we include a dependence on the hidden variable of the translation model in the direct translation model. Therefore, we are able to use statistical alignment models, which have been shown to be a very powerful component for statistical machine translation systems.

In speech recognition, training the parameters of the acoustic model by optimizing the (average) mutual information and conditional entropy as they are defined in information theory is a standard approach (Bahl et al., 1986; Ney, 1995). Combining various probabilistic models for speech and language modeling has been suggested in (Beyerlein, 1997; Peters and Klakow, 1999).

7 Conclusions

We have presented a framework for statistical MT for natural languages, which is more general than the widely used source-channel approach. It allows a baseline MT system to be extended easily by adding new feature functions. We have shown that a baseline statistical MT system can be significantly improved using this framework.

There are two possible interpretations for a statistical MT system structured according to the source-channel approach, hence including a model for $Pr(e_1^I)$ and a model for $Pr(f_1^J|e_1^I)$. We can interpret it as an approximation to the Bayes decision rule in Eq. 2 or as an instance of a direct maximum entropy model with feature functions $\log Pr(e_1^I)$ and $\log Pr(f_1^J|e_1^I)$. As soon as we want to use model scaling factors, we can only do this in a theoretically justified way using the second interpretation. Yet, the main advantage comes from the large number of additional possibilities that we obtain by using the second interpretation.

An important open problem of this approach is the handling of complex features in search. An interesting question is to come up with features that allow an efficient handling using conventional dynamic programming search algorithms.

In addition, it might be promising to optimize the parameters directly with respect to the error rate of the MT system as is suggested in the field of pattern and speech recognition (Juang et al., 1995; Schlüter and Ney, 2001).

References

L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer. 1986. Maximum mutual information estimation of hidden markov model parameters. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 49–52, Tokyo, Japan, April.

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. Beyerlein. 1997. Discriminative model combination. In *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 238–245, Santa Barbara, CA, December.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

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

B. H. Juang, W. Chou, and C. H. Lee. 1995. Statistical and discriminative methods for speech recognition. In A. J. R. Ayuso and J. M. L. Soler, editors, *Speech Recognition and Coding - New Advances and Trends*. Springer Verlag, Berlin, Germany.

H. Ney. 1995. On the probabilistic-interpretation of neural-network classifiers and discriminative training criteria. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(2):107–119, February.

S. Nießen, F. J. Och, G. Leusch, and H. Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proc. of the Second Int. Conf. on Language Resources and Evaluation (LREC)*, pages 39–45, Athens, Greece, May.

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

K. A. Papineni, S. Roukos, and R. T. Ward. 1997. Feature-based language understanding. In *European Conf. on Speech Communication and Technology*, pages 1435–1438, Rhodes, Greece, September.

K. A. Papineni, S. Roukos, and R. T. Ward. 1998. Maximum likelihood and discriminative training of direct translation models. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 189–192, Seattle, WA, May.

K. A. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY, September.

J. Peters and D. Klakow. 1999. Compact maximum entropy language models. In *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, CO, December.

R. Schlüter and H. Ney. 2001. Model-based MCE bound to the true Bayes' error. *IEEE Signal Processing Letters*, 8(5):131–133, May.

W. Wahlster. 1993. Verbmobil: Translation of face-to-face dialogs. In *Proc. of MT Summit IV*, pages 127–135, Kobe, Japan, July.