Word-Level Confidence Estimation for Machine Translation using Phrase-Based Translation Models

Nicola Ueffing and Hermann Ney
Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
{ueffing,ney}@informatik.rwth-aachen.de

Abstract

Confidence measures for machine translation is a method for labeling each word in an automatically generated translation as correct or incorrect. In this paper, we will present a new approach to confidence estimation which has the advantage that it does not rely on system output such as N-best lists or word graphs as many other confidence measures do. It is, thus, applicable to any kind of machine translation system.
Experimental evaluation has been performed on translation of technical manuals in three different language pairs. Results will be presented for different machine translation systems to show that the new approach is independent of the underlying machine translation system which generated the translations. To the best of our knowledge, the performance of the new confidence measure is better than that of any existing confidence measure.

1 Introduction

The work presented in this paper deals with confidence estimation for machine translation (MT). Since sentences produced by a machine translation system are often incorrect but may contain correct parts, a method for identifying those correct parts and finding possible errors is desirable. For this purpose, each word in the generated target sentence is assigned a value expressing the confidence that it is correct.

Confidence measures have been extensively studied for speech recognition, but are not well known in other areas. Only recently have researchers started to investigate confidence measures for machine translation (Blatz et al., 2004; Gandrabur and Foster, 2003; Quirk, 2004; Ueffing et al., 2003).

We apply word confidence measures in MT as follows: For a given translation generated by a machine translation system, we determine a confidence value for each word and compare it to a threshold. All words whose confidence is above this threshold are tagged as correct and all others are tagged as incorrect translations. The threshold is optimized on a distinct development set beforehand.

Possible applications for confidence measures include

- post-editing, where words with low confidence could be marked as potential errors,
- improving translation prediction accuracy in trans-type-style interactive machine translation (Gandrabur and Foster, 2003; Ueffing and Ney, 2005),
- combining output from different machine translation systems: hypotheses with low confidence can be discarded before selecting one of the system translations (Akiba et al., 2004), or the word confidence scores can be used for generating new hypotheses from the output of different systems (Jayaraman and Lavie, 2005), or the sentence confidence value can be employed for re-ranking (Blatz et al., 2003).

In this paper, we will present several approaches to word-level confidence estimation and develop a new phrase-based confidence measure which is independent of the machine translation system which

generated the translation. The paper is organized as follows: In section 2, we will briefly review the statistical approach to machine translation. The phrase-based translation system, which serves as basis for the new confidence measure, will be presented in section 2.2. Section 3 will give an overview of related work on confidence estimation for statistical machine translation (SMT). In section 4, we will describe methods for confidence estimation which make use of SMT system output such as word graphs and N-best lists. In section 5, we will present the new phrase-based confidence measure. Section 6 contains a short description of an IBM-1 based confidence measure to which we will compare the other measures. Experimental evaluation and comparison of the different confidence measures will be shown in section 7, and section 8 will conclude the paper.

2 Statistical machine translation

2.1 General

In statistical machine translation, the translation is modeled as a decision process: Given a source string $f_1^J = f_1 \dots f_j \dots f_J$, we seek the target string $e_1^I = e_1 \dots e_i \dots e_I$ with maximal posterior probability:

$$\begin{aligned} \hat{e}_1^{\hat{I}} &= \underset{I, e_1^I}{\text{argmax}} \left\{ Pr(e_1^I \mid f_1^J) \right\} \qquad (1) \\ &= \underset{I, e_1^I}{\text{argmax}} \left\{ Pr(f_1^J \mid e_1^I) \cdot Pr(e_1^I) \right\} \end{aligned}$$

Through this decomposition of the probability, we obtain two knowledge sources: the translation model $Pr(f_1^J \mid e_1^I)$ and the language model $Pr(e_1^I)$. Both of them can be modeled independently of each other. The translation model is responsible for linking the source string f_1^J and the target string e_1^I, i.e. it captures the semantics of the sentence. The target language model captures the well-formedness or the syntax in the target language. Nowadays, most of the state-of-the-art SMT systems are based on bilingual phrases (Bertoldi et al., 2004; Koehn et al., 2003; Och and Ney, 2004; Tillmann, 2003; Vogel et al., 2004; Zens and Ney, 2004). Note that those phrases are sequences of words in the two languages and not necessarily phrases in the linguistic sense. A more detailed description of a phrase-based approach to statistical machine translation will be given in section 2.2.

2.2 Review of phrase-based translation system

For the confidence measures which will be introduced in section 5, we use a state-of-the-art phrase-based approach as described in (Zens and Ney, 2004). The key elements of this translation approach are bilingual phrases, i.e. pairs of source and target language phrases where a phrase is simply a contiguous sequence of words. These bilingual phrases are extracted from a word-aligned bilingual training corpus.

We will present the equations for a monotone search here in order to keep the equations simple. Let (j_0^K, i_0^K) be a segmentation of the source sentence into phrases, with the corresponding (bilingual) phrase pairs $(\tilde{f}_k, \tilde{e}_k) = (f_{j_{k-1}+1}^{j_k}, e_{i_{k-1}+1}^{i_k}), k = 1, \dots, K$. The phrase-based approach to SMT is then expressed by the following equation:

$$\begin{aligned} \hat{e}_1^{\hat{I}} = \underset{j_0^K, i_0^K, I, e_1^I}{\text{argmax}} \Bigg\{ & \prod_{i=1}^{I} \left[c_1 \cdot p(e_i \mid e_{i-2}^{i-1})^{\lambda_1} \right] \qquad (2) \\ & \cdot \prod_{k=1}^{K} \Big[c_2 \cdot \; p(\tilde{f}_k \mid \tilde{e}_k)^{\lambda_2} \cdot p(\tilde{e}_k \mid \tilde{f}_k)^{\lambda_3} \\ & \cdot \prod_{j=j_{k-1}+1}^{j_k} p(f_j \mid \tilde{e}_k)^{\lambda_4} \cdot \prod_{i=i_{k-1}+1}^{i_k} p(e_i \mid \tilde{f}_k)^{\lambda_5} \Big] \Bigg\} , \end{aligned}$$

where $p(\tilde{f}_k \mid \tilde{e}_k)$ and $p(\tilde{e}_k \mid \tilde{f}_k)$ are the phrase lexicon models in both translation directions. The phrase translation probabilities are computed as a log-linear interpolation of the relative frequencies and the IBM-1 probability. The single word based lexicon models are denoted as $p(f_j \mid \tilde{e}_k)$ and $p(e_i \mid \tilde{f}_k)$, respectively. $p(f_j \mid \tilde{e}_k)$ is defined as the IBM-1 model probability of f_j over the whole phrase $\tilde{e}_k$, and $p(e_i \mid \tilde{f}_k)$ is the inverse model, respectively.

c_1 is the so-called word penalty, and c_2 is the phrase penalty, assigning constant costs to each target language word/phrase. The language model is a trigram model with modified Kneser-Ney discounting and interpolation (Stolcke, 2002). The search determines the target sentence and segmentation which maximize the objective function.

As equation 2 shows, the sub-models are combined via weighted log-linear interpolation. The model scaling factors $\lambda_1, \dots, \lambda_5$ and the word and phrase penalties are optimized with respect to some evaluation criterion (Och, 2003), e.g. BLEU score.

3 Confidence measures for SMT

3.1 Related work

In this paper, we will present a new approach to word-level confidence estimation which makes explicit use of a phrase-based translation model. Most of the word-level confidence measures which have been presented in the literature so far are either based on relatively simple translation models such as IBM-1 (Blatz et al., 2003) or make use of information provided by the SMT system such as N-best lists or word graphs (Blatz et al., 2003; Gandrabur and Foster, 2003; Ueffing et al., 2003). In contrast to this, our method is based on a state-of-the-art statistical machine translation model, but nevertheless is independent of the machine translation system which generates the translation hypotheses.

The word-level confidence measures which showed the best performance in comparative experiments (Blatz et al., 2003) are word posterior probabilities and the IBM-1 based measure. Our new confidence measure will be compared to those approaches in section 7.3.

3.2 Word posterior probabilities

The confidence of a target word can be expressed by its posterior probability, i.e. the probability of the word to occur in the target sentence, given the source sentence. Consider a target word e occurring in the sentence in position i[1]. The posterior probability of this event can be determined by summing over all possible target sentences e_1^I containing the word e in position i:

$$p_i(e, f_1^J) \quad = \quad \sum_{I, e_1^I:\, e_i = e} p(e_1^I, f_1^J) \qquad (3)$$

This value has to be normalized in order to obtain a probability distribution over all possible target words:

$$p_i(e \,|\, f_1^J) \quad = \quad \frac{p_i(e, f_1^J)}{\sum_{e'} p_i(e', f_1^J)} \qquad (4)$$

[1]This is a rather strict assumption, because the position of a word in the target sentence can differ largely due to reorderings in the translation process. We present this variant here to keep the notation simple. Improved methods will be shown in the following sections.

4 System based confidence measures

In this section, we will present confidence measures which are based on N-best lists or word graphs generated by the SMT system. Those are representations of the space of the most likely translations of the source sentence.

The summation given in equation 3 is performed over all sentences which are contained in the N-best list or word graph. For a more detailed description, see (Ueffing et al., 2003).

4.1 Word graph based approach

The word posterior probability $p_i(e \,|\, f_1^J)$ can be calculated over a word graph using the forward-backward algorithm.

Let n', n be nodes in a word graph, and (n', n) the directed edge connecting them. The edge is annotated with a target word which we denote by $e(n', n)$ and the probability which this word contributes to the overall sentence probability, denoted by $p(n', n)$.

The forward probability $\Phi_i(n', n)$ of an edge is the probability of reaching this edge from the source of the graph, where the word $e(n', n)$ is the i-th word on the path. It can be obtained by summing the probabilities of all incoming paths of length $i - 1$, which allows for recursive calculation. This leads to the following formula:

$$\Phi_i(n', n) \quad = \quad p(n', n) \cdot \sum_{n''} \Phi_{i-1}(n'', n') \; .$$

The backward probability expresses the probability of completing a sentence from the current edge, i.e. of reaching the sink of the graph. It can be determined recursively in descending order of i as follows:

$$\Psi_i(n', n) \quad = \quad p(n', n) \cdot \sum_{n^*} \Psi_{i+1}(n, n^*) \; .$$

Using the forward-backward algorithm, the word posterior probability of word e in position i is determined by combining the forward and backward probabilities of all edges which are annotated with e. This yields

$$p_i(e, f_1^J) = \sum_{(n', n):\, e(n', n) = e} \frac{\Phi_i(n', n) \cdot \Psi_i(n', n)}{p(n', n)} \; . \qquad (5)$$

Note that (for computational reasons) the term $p(n', n)$ is included both in the forward and in the

backward probability so that we have to divide the product by this term.

To obtain a posterior probability, a normalization, as shown in equation 4, has to be performed. The normalization term $\alpha := \sum_{e'} p_i(e', f_1^J)$ corresponds to the probability mass contained in the word graph and can be calculated by summing the backward probabilities of all outgoing edges leaving the source s of the graph:

$$\alpha = \sum_{(s,n)} \Psi_1(s,n) \; .$$

As stated above, the position of word e in the target sentence can vary due to reorderings in the translation process. Therefore, we would like to relax the condition that e has to occur exactly in position i. This can be achieved by introducing a window of size t over the neighboring target positions and computing the sum of the word posterior probabilities over all positions $i-t, \ldots, i, \ldots, i+t$. In our experiments we found that a window over ± 3 positions yields the best performance.

4.2 N-best list based approach

N-best lists are an alternative representation of the space of translation hypotheses. They have the advantage that the Levenshtein alignment between a hypothesis and all sentences contained in the list can be performed easily. This makes it possible to consider not only target sentences, which contain the word e exactly in a position i (as given in equation 3), but to allow for some variation.

Let $\mathcal{L}(e_1^I, \tilde{e}_1^{\tilde{I}})$ be the Levenshtein alignment between sentences e_1^I and $\tilde{e}_1^{\tilde{I}}$. Then, $\mathcal{L}_i(e_1^I, \tilde{e}_1^{\tilde{I}})$ denotes the Levenshtein alignment of word e_i, i.e. the word in sentence $\tilde{e}_1^{\tilde{I}}$ which e_i is Levenshtein-aligned to.

The word posterior probability is then calculated by summing over all target sentences containing word e in a position which is Levenshtein-aligned to i:

$$p_i(e|f_1^J, I, e_1^I) = \frac{p_i(e, f_1^J, I, e_1^I)}{\sum_{e'} p_i(e', f_1^J, I, e_1^I)} \; ,$$

where

$$p_i(e, f_1^J, I, e_1^I) = \sum_{\tilde{I}, \tilde{e}_1^{\tilde{I}}:\, \mathcal{L}_i(e_1^I, \tilde{e}_1^{\tilde{I}}) = e} p(\tilde{e}_1^{\tilde{I}}, f_1^J) \; . \quad (6)$$

The confidence of word e then depends on the source sentence f_1^J as well as the target sentence e_1^I, because the whole target sentence is relevant for the Levenshtein alignment.

5 Phrase-based confidence measures

In contrast to the approaches presented in section 4, the phrase-based confidence measures do not not use the context information at the sentence level, but only at the phrase level. We want to determine a sort of marginal probability $Q(e, f_1^J)$. Therefore, we extract all source phrases f_j^{j+s} which occur in the given source sentence. For such source phrases, we find the possible translations e_i^{i+t} in the bilingual phrase lexicon. The confidence of target word e is then calculated by summing over all phrase pairs (f_j^{j+s}, e_i^{i+t}) where the target part e_i^{i+t} contains the word e.

Let $p(e_i^{i+t})$ be the language model score of the target phrase together with the word penalty c_1, i.e.

$$p(e_i^{i+t}) = \prod_{i'=i}^{i+t} c_1 \cdot p(e_{i'} \,|\, e_{i'-2}^{i'-1})^{\lambda_1} \; .$$

Analogously, define $p(f_j^{j+s}, e_i^{i+t})$ as the score of the phrase pair which consists of the phrase penalty and the phrase and word lexicon model scores (cf. section 2.2). Following equation 2, the (unnormalized) confidence is then determined as:

$$Q(e, f_1^J) = \sum_{j=1}^{J} \sum_{s=0}^{\min\{s_{\max}, J-j\}} \sum_{e_i^{i+t}:\, e \in e_i^{i+t}} p(e_i^{i+t}) \cdot p(f_j^{j+s}, e_i^{i+t}) \; , \quad (7)$$

where $s \leq s_{\max}$ and t are source and target phrase lengths, $s_{\max}$ being the maximal source phrase length.

In equation 7, the language model only determines the probability of the words within the target part of the phrase, and not across the phrase boundaries, because we consider only the single target phrases without context. Therefore, we assumed that the language model would not have much influence on the confidence estimation and also investigated a model without a language model. The same holds for word and phrase penalty: In the translation process they are useful for adjusting the length of the

generated target hypothesis and for assigning more weight to longer phrases. Since this does not make much sense in our setting, we also investigated confidence estimation without word and phrase penalty.

Note that the value calculated in equation 7 is not normalized. In order to obtain a word posterior probability, we divide this value by the sum over the (unnormalized) confidence of all target words:

$$p_{phr}(e \mid f_1^J) = \frac{Q(e, f_1^J)}{\sum_{e'} Q(e', f_1^J)} \ . \tag{8}$$

Unlike the word posterior probabilities presented in the previous section, this value is completely independent of the target sentence position in which the word e occurs.

As stated in section 2.2, the scaling factors of the different sub-models and the penalties in the translation system are optimized with respect to some evaluation criterion. But since the values which are optimal for translation are not necessarily optimal for confidence estimation, we perform optimization here as well: We train the probability models on the training corpus, estimate the word confidences on the development corpus, and optimize the scaling factors with respect to the classification error rate described in section 7.2. The optimization is performed with the Downhill Simplex algorithm (Press et al., 2002).

6 IBM-1 based approach

Another type of confidence measure which does not rely on system output and is thus applicable to any kind of machine translation system is the IBM-1 model based confidence measure which was introduced in (Blatz et al., 2003). We modified this confidence measure because we found that the average lexicon probability used there is dominated by the maximum. Therefore, we determine the *maximal* translation probability of the target word e over the source sentence words:

$$p_{\mathrm{IBM-1}}(e|f_1^J) = \max_{j=0,\ldots,J} p(e|f_j) \ , \tag{9}$$

where f_0 is the "empty" source word (Brown et al., 1993). The probabilities $p(e|f_j)$ are word-based lexicon probabilities.

Investigations on the use of the IBM-1 model for word confidence measures showed promising results (Blatz et al., 2003; Blatz et al., 2004). Thus, we apply this method here in order to compare it to the other types of confidence measures.

7 Experiments

7.1 Experimental setting

The experiments were performed on three different language pairs. All corpora were compiled in the EU project TransType2; they consist of technical manuals. The corpus statistics are given in table 1. The SMT systems that the confidence estimation was performed for were trained on these corpora. The same holds for the probability models that were used to estimate the word confidences.

We used several (S)MT systems for testing the confidence measures. A detailed analysis will be given for two of them; the so-called alignment template system (Och and Ney, 2004), (denoted as AT in the tables) and the phrase-based translation system described in section 2.2 (denoted as PBT in the tables). They are both state-of-the-art SMT systems. We produced single best translations, word graphs and N-best lists on all three language pairs using these systems. The translation quality in terms of WER, PER (position independent word error rate), BLEU and NIST score is given in tables 2 and 3. We see that the best results are obtained on Spanish to English translation, followed by French to English and German to English.

Two more translation systems were used for comparative experiments: One is a statistical MT system which is based on a finite state architecture (FSA). For a description of this system, see (Kanthak et al., 2005). Additionally, we used translations generated by Systran[2]. Table 3 presents the translation error rates and scores for all systems on the German → English test corpus. These hypotheses were used to investigate whether the phrase-based confidence measures perform well independently of the translation system.

All three SMT systems (AT, PBT and FSA) show very similar performance on the German → English test corpus. The fact that Systran generates translations of much lower quality is due to the fact that the technical manuals are very specific in terminology, and the SMT systems have been trained on similar corpora.

[2] http://babelfish.altavista.com/tr, June 2005

Table 1: Statistics of the training, development and test corpora.

		French	English	Spanish	English	German	English
TRAIN	Sentences	53 046		55 761		49 376	
	Running Words	680 796	628 329	752 606	665 399	537 464	589 531
	Vocabulary	15 632	13 816	11 050	7 956	23 845	13 223
DEV	Sentences	994		1 012		964	
	Running Words	11 674	10 903	15 957	14 278	10 462	10 642
	OOVs	184	141	54	27	147	29
TEST	Sentences	984		1 125		996	
	Running Words	11 709	11 177	10 106	8 370	11 704	12 298
	OOVs	204	201	69	49	485	141

Table 2: Translation quality of systems AT and PBT on the test corpora described in table 1.

	AT		PBT	
	S→E	F→E	S→E	F→E
WER[%]	29.6	54.8	26.1	54.9
PER[%]	20.1	43.7	17.5	43.4
BLEU[%]	63.4	31.5	66.9	31.3
NIST	8.80	6.64	8.98	6.62

Table 3: Translation quality of all MT systems on the German → English test corpus.

	AT	PBT	FSA	Systran
WER[%]	62.7	61.6	63.2	79.2
PER[%]	49.8	49.6	50.4	66.4
BLEU[%]	26.6	25.7	26.5	12.0
NIST	5.92	5.72	5.79	4.09

To determine the true class of each word in a generated translation hypothesis, we use the word error rate (WER). That is, a target word is considered correct if it is aligned to itself in the Levenshtein alignment between hypothesis and reference translation(s). We also investigated PER based classification, but since the tendencies of the results were similar, we omit them here.

7.2 Evaluation metrics

After computing the confidence measure, each generated word is tagged as either *correct* or *false*, depending on whether its confidence exceeds the tagging threshold that has been optimized on the development set beforehand. The performance of the confidence measure is evaluated using the **C**lassification **E**rror **R**ate (CER). This is defined as the number of incorrect tags divided by the total number of generated words in the translated sentence. The baseline CER is determined by assigning the most frequent class to all translations. In the case that the most frequent class is "correct" (meaning at least half of the words in the generated translation are correct w.r.t. to WER), this is the number of substitutions and insertions, divided by the number of generated words. The CER strongly depends on the tagging threshold. Therefore, the tagging threshold is adjusted beforehand (to minimize CER) on a development corpus *distinct* to the test set.

7.3 Experimental results

Table 4 shows the performance of all different confidence measures on the hypotheses generated by the alignment template system and the phrase-based system. For the baseline CER, we determined the 90%- and 99%-confidence intervals using the bootstrap estimation method described in (Bisani and Ney, 2004)[3]. We see that, in all settings but one, the word graph and the N-best list based method outperform the IBM-1 based confidence measure. On French → English, the improvement over the baseline is significant at the 1%-level for these methods, whereas on Spanish → English this is only the case at 10%. The performance of the N-best list based approach is better than that of the word graph based

[3]The tool is freely available from http://www-i6.informatik.rwth-aachen.de/web/Software/index.html

confidence measures for the alignment template system. This is probably due to the fact that the former can take the Levenshtein alignment into account and thus estimate the word confidence more reliably.

The phrase-based confidence measures show a performance which is clearly better than that of the other methods. We obtain a relative improvement of up to 7.8% over the best existing method on these language pairs. The improvement over the baseline is significant even at the 1%-level in all cases.

When analyzing the impact of the different submodels in the phrase-based approach, we found that the language model does not have much impact on the confidence estimation. There are only slight variations in the CER if the model is omitted. The word and phrase penalty on the other hand seem to be important (with one exception in the first setting).

The evaluation of the system-independent confidence measures (i.e. those based on IBM-1 and the new phrase-based method we presented) for four different translation systems is shown in table 5. We see that, for all of them, the phrase-based approach outperforms the IBM-1 based method significantly. The largest gain in terms of CER is achieved for the Systran translations: 23.8% relative over the IBM-1 based measure.

8 Conclusion and outlook

We presented a new approach to word-level confidence estimation for machine translation which makes use of bilingual phrases. By using models from a state-of-the-art phrase-based statistical machine translation system, the word confidences are estimated only on the basis of single best system output. Unlike other confidence measures, this does not rely on information from the machine translation system which generated the translation.

Experimental evaluation on three different language pairs and on output from structurally different translation systems showed that the new confidence measures perform better than existing confidence measures in all cases. The application on output from different MT systems yielded a significant reduction of the error rate over the existing measures. This proves that the method is well-suited for word confidence estimation on statistical as well as non-statistical MT systems.

The task investigated in this work was a text translation task in the domain of technical manuals. We are currently investigating the use of word-level confidence measures on data from the European parliament. It will be interesting to see whether a similar performance can be achieved on this large vocabulary speech translation task.

Acknowledgement

This work was partly funded by the European Union under the RTD project TransType2 (IST–2001–32091), and under the integrated project TC-STAR – Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738).

References

Y. Akiba, E. Sumita, H. Nakaiwa, S. Yamamoto, and H. G. Okuno. 2004. Using a mixture of n-best lists from multiple MT systems in rank-sum-based confidence measure for MT outputs. In *Proc. CoLing*, pages 322–328, August.

N. Bertoldi, R. Cattoni, M. Cettolo, and M. Federico. 2004. The ITC-irst statistical machine translation system for IWSLT-2004. In *Proc. IWSLT*, pages 51–58, Kyoto, Japan, September.

M. Bisani and H. Ney. 2004. Bootstrap estimates for confidence intervals in asr performance evaluationx. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 409–412, Montreal, Canada, May.

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2003. Confidence estimation for machine translation. Final report, JHU/CLSP Summer Workshop. `http://www.clsp.jhu.edu/ws2003/groups/estimate/`.

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2004. Confidence estimation for machine translation. In *Proc. CoLing*, pages 315–321, Geneva, Switzerland, August.

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, June.

S. Gandrabur and G. Foster. 2003. Confidence estimation for text prediction. In *Proc. CoNLL*, pages 95–102, Edmonton, Canada, May.

S. Jayaraman and A. Lavie. 2005. Multi-engine machine translation guided by explicit word matching.

Table 4: CER for different confidence measures, reference based on WER. Hypotheses from the alignment template system and the phrase-based system. The best value is printed in bold.

	alignment template system		phrase-based system	
Model	S → E	F → E	S → E	F → E
Baseline	20.8	42.5	19.2	42.7
99%-confidence interval	[18.8,22.7]	[40.1,44.7]	[17.2,21.2]	[40.4,45.0]
90%-confidence interval	[19.6,22.1]	[40.9,43.9]	[17.9,20.5]	[41.2,44.2]
Word graphs from the system (eq. 5)	20.1	32.9	17.9	30.5
N-best lists from the system (eq. 6)	19.8	31.9	17.9	30.9
phrase-based (eq. 8)	17.5	**30.2**	**16.5**	**30.0**
without language model	**17.4**	30.3	**16.5**	30.3
without word and phrase penalty	17.5	30.6	16.9	30.5
IBM-1 (eq. 9)	20.0	34.1	18.3	35.1

Table 5: CER for different confidence measures on the German → English test set, reference based on WER. Hypotheses from different MT systems.

Model	AT	PBT	FSA	Systran
Baseline	49.2	48.4	46.6	37.4
99%-confidence interval	[48.6,53.1]	[45.9,50.7]	[44.2,49.0]	[36.0,38.9]
phrase-based (eq. 8)	27.6	26.4	30.2	24.3
IBM-1 (eq. 9)	32.8	32.8	37.0	31.9

In *Proc. EAMT*, pages 143–152, Budapest, Hungary, May.

S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney. 2005. Novel reordering approaches in phrase-based statistical machine translation. In *ACL 2005 – Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 167–174, Ann Arbor, Michigan, June.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. HLT-NAACL*, pages 127–133, Edmonton, Canada, May/June.

F. J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, December.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. ACL*, pages 160–167, Sapporo, Japan, July.

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

C. Quirk. 2004. Training a sentence-level machine translation confidence metric. In *Proc. LREC*, pages 825–828, Lisbon, Portugal, May.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. ICSLP*, volume 2, pages 901–904, Denver.

C. Tillmann. 2003. A projection extension algorithm for statistical machine translation. In *Proc. EMNLP*, pages 1–8, Sapporo, Japan, July.

N. Ueffing and H. Ney. 2005. Application of word-level confidence measures in interactive statistical machine translation. In *Proc. EAMT*, pages 262–270, Budapest, Hungary, May.

N. Ueffing, K. Macherey, and H. Ney. 2003. Confidence Measures for Statistical Machine Translation. In *Proc. MT Summit IX*, pages 394–401, New Orleans, LA, September.

S. Vogel, S. Hewavitharana, M. Kolss, and A. Waibel. 2004. The ISL statistical translation system for spoken language translation. In *Proc. IWSLT*, pages 65–72, Kyoto, Japan, September.

R. Zens and H. Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proc. HLT-NAACL*, pages 257–264, Boston, MA, May.