Application of Word-Level Confidence Measures in Interactive Statistical Machine Translation

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. In this paper, we will address the question of how to efficiently integrate word confidence measures into a state-of-the-art interactive statistical machine translation system and improve prediction performance. Different methods will be presented: the selection of words according to their confidence as well as the rejection which has not been investigated so far. Experimental evaluation with respect to prediction accuracy of the system and typing effort saved by a user will show the improved prediction quality. Additionally, we will describe novel methods exploiting knowledge about a correct prefix of the target sentence for confidence estimation. These further increase the interactive system's performance.

1 Introduction

The work presented in this paper deals with the application of confidence estimation in an interactive machine translation system. This system aims at improving the productivity of human translators by suggesting translations of the source text and taking text into account that the user has typed already.
We integrated confidence estimation into such an interactive system in order to improve the quality of translations predicted by the system. Since the goal is to reduce users' effort, one has to consider the gain in keystrokes needed to type the translation as well as the time that he or she spends on reading and deciding whether to accept a suggestion. That is, the system has to keep the balance between the benefit of long predictions and the negative effect of incorrect predictions. We apply confidence estimation as a way to achieve this balance.
The system uses confidence estimation in two different ways: for the *rejection* of words with low confidence and for the *selection* of words for extension based on their confidence. Those methods improve the prediction accuracy of the system and reduce the typing effort of a human user. We further improve these methods by taking a correct prefix that the user has already entered into account. The confidence estimation can be modified to account only for untranslated words in the source sentence.

2 Interactive Statistical Machine Translation

In statistical machine translation (SMT), the translation is modeled as a decision process: For a given source string $f_1^J = f_1 \ldots f_j \ldots f_J$, we seek for the target string $e_1^I = e_1 \ldots e_i \ldots e_I$ with maximal posterior probability:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I \mid f_1^J) \right\}$$

In the setting that we investigate in this paper, a state-of-the-art SMT system is employed in an interactive translation system in the following way: For a given source sentence, the system generates a translation. A human translator checks this translation from left to right, correcting the first error. The SMT system then proposes a new extension, taking the correct prefix $e_1^i = e_1 \ldots e_i$ into account. These steps are repeated until the whole input sentence has been correctly translated. In the resulting decision rule, we maximize over all possible extensions e_{i+1}^I of e_1^i:

$$\hat{e}_{i+1}^{\hat{I}} = \underset{I, e_{i+1}^I}{\operatorname{argmax}} \left\{ Pr(e_{i+1}^I \mid e_1^i,\, f_1^J) \right\}$$

For reasons of simplicity, this equation is formulated on the word level. In the actual implementation, the same method is applied on the character level, and the search for the extension is per-

formed after each keystroke of the human translator. Note that this decision can be taken on the word or character level even though the translation system makes use of bilingual phrases.
This approach requires a highly efficient search, because human users will only accept response times of fractions of a second. To achieve this, the SMT system computes a word graph representing a subset of the possible translations of the input sentence (Ueffing et al., 2002). This representation of the search space is then used for the efficient computation of the extension. For a description of the interactive SMT system, see (Och et al., 2003).

The concept of *interactive machine translation* was first suggested by (Foster et al., 1996). The basic idea is to provide an environment to a human translator that interactively reacts upon the input as the user writes or corrects the translation. In such an approach, the system suggests an extension of a sentence that the human user either accepts or ignores. An implementation of such a tool was performed in the TransType project (Foster et al., 1996; Foster et al., 1997; Langlais et al., 2000) and further improved within TransType2 (Atos Origin Spain et al., 2002; Civera et al., 2004).

3 Confidence Measures for SMT

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., 2003; Gandrabur and Foster, 2003; Blatz et al., 2004; Quirk, 2004; Ueffing and Ney, 2004).
For the application in an interactive environment, we need confidence measures that operate on the word level (instead of the sentence level) and that can be computed very efficiently.
(Gandrabur and Foster, 2003) studies the application of word-level confidence measures for translation prediction in an interactive SMT system. The SMT system applied there is a rather simple model that combines a trigram language model and the IBM translation model 2 (Brown et al., 1993). In contrast to this, we apply a fully-fledged SMT system on the basis of bilingual phrases (Och and Ney, 2004; Bender et al., 2004), taking a large number of different submodels into account. Moreover, the system described in (Gandrabur and Foster, 2003) predicts extensions of up to four words, whereas our system translates the whole input sentence. In addition to the prediction of words based on the confidence as proposed in (Gandrabur and Foster, 2003), we also study a new approach of rejecting words with low confidence. Furthermore, we will introduce a method to make use of a given prefix (which is known to be correct) in the confidence estimation.

4 Application in Interactive MT

If the human user has entered a character or accepted a part of the translation proposed by the system, the interactive SMT system starts searching for an appropriate extension of this prefix. It performs the following steps (cf. (Och et al., 2003)):

1. locate the prefix in the word graph
2. search for the best extension in the word graph
3. if no good extension is found: use the language model for prediction

Word confidence measures have been integrated into the interactive SMT system in steps 2 and 3. We investigated different methods of incorporating them which will be described in sections 4.1 through 4.3.

For the application in the interactive system, we implemented a word confidence measure based on the IBM translation model 1 (Brown et al., 1993), similar to the one described in (Blatz et al., 2004). We chose this because it relies only on the source sentence and the proposed extension, and not on an N-best list or an additional confidence estimation layer as many other word confidence measures do. Thus, it can be calculated very fast during search. Moreover, its performance in identifying correct words is similar to that of other word confidence measures as the results presented in (Blatz et al., 2003; Blatz et al., 2004; Ueffing and Ney, 2004) show. However, we modified this confidence measure by replacing the *average* by the *maximal* lexicon probability, because we found that the average is dominated by this maximum. The word confidence $c(e)$ is then given by

$$c(e) = \max_j p(e \mid f_j) \; . \qquad (1)$$

Informal experiments showed that this improves performance compared to the original variant.

4.1 Selection of Words

One way of incorporating confidence measures into the interactive system is to choose the proposed word according to the confidence. This approach is pursued in (Gandrabur and Foster, 2003). Since the SMT system applied in our work is more sophisticated than the one investigated there, we expect the gain from this selection to be lower in our case.
We investigated the confidence based prediction in the interactive SMT system in search steps 2 and 3 explained above. The confidence of the word and its original score assigned by the SMT system are combined in a log-linear manner.

4.2 Rejection of Words

A novel way of incorporating confidence estimation into the interactive system is to reject proposed words if their confidence is below a given threshold. When the SMT system searches for an extension of a given prefix, we calculate the confidence for each word in the possible extensions. For the words contained in the word graph, this calculation can be done already when the word graph is constructed.
If at a certain point all words that are possible extensions are rejected, the system will stop predicting translations. Thus, the system does not necessarily propose whole-sentence extensions anymore. This approach aims at preventing the prediction of long, incorrect extensions. Once the user has entered another character, the interactive system searches for an extension again.

4.3 Use of Prefix Information

The confidence estimation described so far does not take into account that in interactive use, the user has accepted and or typed a part of the translation already. This prefix is known to be correct, and we will introduce a way of exploiting this knowledge in this chapter.
Since the confidence of target word e is calculated as the maximal lexicon probability over the source sentence, we can restrict this calculation to those source words that are not covered by the given prefix. An example of this will be given in table 5 in section 5.4. If a prefix word has no correspondence in the source sentence, this will not influence the confidence estimation. Let e_1^i a given prefix covering the source words $F(e_1^i)$. The confidence estimation introduced in equation 1 is then modified as follows:

$$c(e) = \max_{f_j \notin F(e_1^i)} p(e \mid f_j) \; . \qquad (2)$$

This prevents the system from proposing translations of source words that have already been translated.

Another way of using the knowledge contained in the prefix is the adaptation of the confidence threshold. Since the prefix words are known to be correct, we assume that they should be accepted by the confidence module. Different source sentences might have different inherent translation difficulties. The fact that the confidence values depend on the source sentence can result in different ranges for those values. To account for this, we compare the confidence of each word in the prefix to the confidence threshold and lower the latter if necessary. In order not to adapt the threshold to outliers, we never lower the threshold by more than half its value.

5 Experimental Results

5.1 Experimental Setting

The experiments were performed on two corpora consisting of technical manuals. The translation directions are French→English and English→German; see table 1 for the corpus statistics. These corpora were compiled within the European project TransType2 (Atos Origin Spain et al., 2002). The SMT systems were trained on the training corpora described in table 1. The lexicon that was used to estimate the word confidences was trained on the same corpora.
The translation prediction experiments were performed using the SMT engine to translate the test corpora and simulating the interactive mode. We will summarize this simulation mode here, for a more detailed description see (Och et al., 2003). Every time the system proposes an extension, this is compared to the reference sentence. The comparison is done from left to right, and that part which matches exactly is accepted by the simulated user. The rest of the extension is discarded, and the system starts proposing new completions, taking the correct prefix into account. This reflects the application, where we attempt to match what a human

user has in mind, and not simply to produce any correct translation.

5.2 Evaluation Metrics

In the experiments reported on this task so far (Gandrabur and Foster, 2003; Och et al., 2003; Civera et al., 2004), the evaluation was performed based on the number of keystrokes or the time saved by a user when typing the reference translation. We account for this by measuring the precision and the recall of the predictions. The recall can be measured by the so-called *keystroke ratio (KSR)* introduced in (Och et al., 2003). It divides the number of keystrokes needed to produce the single reference translation using the interactive translation system by the number of keystrokes needed to type the reference translation. Hence, a keystroke ratio of 1 means that the system was never able to suggest a correct extension.

But this error metric has the shortcoming that it does not penalize long predictions of bad quality, e.g. the prediction of 10 incorrect words results in the same KSR as the prediction of one incorrect word. It can be seen as a metric measuring **recall** error on the proposed characters versus the reference.

To overcome this problem, we modify the metric by introducing a term that determines the ratio of proposed characters in the proposed extension that are correct, i.e. we model **precision** of the predictions as well. This accounts also for the reading time that a user spends on predictions even if he or she does not accept them.

The combination of precision and recall using the harmonic mean yields the **prediction F-measure**

$$F_{pred} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

where *Recall* $= 1-$KSR. This evaluation metric lies in the interval between 0 and 1 and measures quality. This metric will penalize a system that proposes very short extensions as well as one that proposes long translations of bad quality.

As additional performance measure, we count the **number of extensions** proposed by the system, i.e. the number of user–system interactions. Every time the system proposes an extension, the user has to read it and to decide whether to accept or not. Thus, a high number of user–system interactions significantly increases the cognitive load of the user. Furthermore, it shows that the quality of predictions is low, because the user often discards them.

5.3 Effect of Confidence-Based Rejection and Selection of Words

When applying the confidence based selection and rejection of words, experiments showed that the system performs best if we base the selection of words in step 2 (described in section 4) only on the score assigned by the SMT system, and consider both the confidence and the SMT system score of the word in step 3. The rejection of words is performed in both step 2 and 3 of the search.

Figure 1 shows the effect of confidence based selection and rejection of words on precision, recall and F_{pred} for different values of the confidence scaling factor and the confidence threshold. A scaling factor of 0 corresponds to not using the confidence estimation for selection of words, and a confidence threshold of 0 corresponds to not rejecting any words.

We ran experiments with several non-zero confidence scaling factors. The best results were obtained for a scaling factor of 0.8, but the prediction performance was very similar for all scaling factors between 0.5 and 1. Thus, we will only present results for the optimal scaling factor here.

As we see in figure 1, recall decreases as the confidence threshold is increased – which is to be expected because the predictions proposed by the system get shorter and sometimes correct words are discarded. On the other hand, precision rises substantially.

When comparing the systems with confidence scaling factors 0 and 0.8, we see that both recall and precision are higher if the selection of words is based on their confidence. This causes the harmonic mean F_{pred} to increase significantly over the baseline. The best result with $F_{pred}= 68.9\%$ is obtained for a confidence scaling factor of 0.8 and a threshold value of 0.5.

5.3.1 Example

Table 2 gives an example of a sentence from the French→English test corpus that is translated in simulated interactive mode with and without the use

Table 1. Statistics of the French–English and training English–German and test sets.

		French	English	English	German
Train:	Sentences	53 046		49 376	
	Running Words	680 796	628 329	589 531	537 464
	Running Words without Punctuation Marks	627 027	573 912	509 902	443 547
	Vocabulary	15 632	13 816	13 223	23 845
	Singletons	4 789	4 032	3 681	9 443
Test:	Sentences	984		996	
	Running Words	11 709	11 177	10 792	9 826
	Running Words without Punctuation Marks	10 889	10 358	10 542	9 595
	OOVs	204	201	1 407	1 931

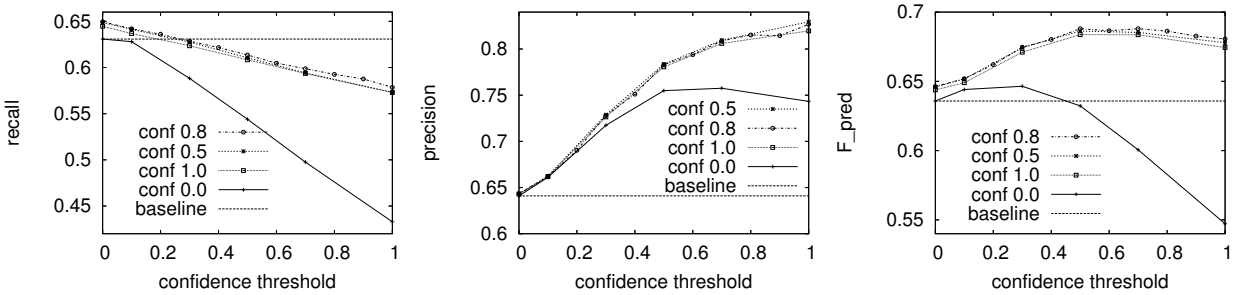

Figure 1. Effect of combination of selection and rejection of words on recall, precision and F_{pred} for different threshold values on the French→English test set. The values of the confidence scaling factor are 0 and 0.8.

of confidence based rejection. The value of the confidence scaling factor is set to 0.8 which is optimal with respect to F_{pred} as the results presented in figure 1 show.

The part of the translation that has been accepted by the user – together with the following character that he/she has entered – is taken as prefix for the new search. The upper part of the table shows the steps that are necessary if none of the words proposed by the SMT engine is discarded, and the lower part shows the translations that are proposed when the confidence threshold is set to 0.5. We see that the correct result is obtained much faster with confidence based rejection than without: only four steps of user–system interaction are necessary instead of seven. Furthermore, bad translations like the word "Remote" and "for" are discarded and not proposed to the user at all.

The improvement of the system's performance is also quite clear in terms of precision and recall: The number of keystrokes necessary to type the reference decreases from 11 to 5; and precision rises drastically from 55.0% to 82.9%. So the gain for the user is substantial; and quality of the proposed translations increases noteworthily.

5.3.2 Extension Lengths and Number of Extensions

As this example shows, the extensions proposed by the system get more accurate, but also significantly shorter. A detailed analysis of this effect is given in table 3. It contains the average length (in characters) of the proposed extensions per source sentence and per extension on the test set. We see that this length drops significantly as more and more translations are discarded. In order to obtain predictions that are not too short, the value for the confidence threshold should not be set too high.

For a new domain or a new language pair, it might not always be clear what a good choice for the confidence threshold will be. One way to account for this is the adaptation of the threshold as described in section 4.3.

The last row of table 3 shows the average number of proposed extensions per source sentence. We see that for thresholds below 0.3, the number of

Table 2. Effect of the combination of rejection and selection on the proposed extensions for different confidence threshold values. Example from the French→English test corpus; simulated interactive mode. The confidence scaling factor is set to 0.8 (optimal w.r.t. F_{pred}).

source	Utilitaire d'impression pour Macintosh vi
reference	Macintosh Printer Utility vi
confidence threshold 0 prefix	
extension	Remote Print for Macintosh vi
prefix	M
extension	anagement print for Macintosh vi
prefix	Mac
extension	intosh vi utility to print
prefix	Macintosh P
extension	rint utility to vi
prefix	Macintosh Printe
extension	r
prefix	Macintosh Printer U
extension	tility to vi
prefix	Macintosh Printer Utility v
extension	i to
confidence threshold 0.5 prefix	
extension	Macintosh vi to
prefix	Macintosh P
extension	rinter
prefix	Macintosh Printer U
extension	tility
prefix	Macintosh Printer Utility v
extension	i to

Table 3. Average extension length (in characters) per extension and number of proposed extensions for different confidence thresholds on the French→English test set. The confidence scaling factor is set to 0.8 (optimal w.r.t. F_{pred}).

confidence threshold	baseline	0.0	0.1	0.2	0.3	0.5	0.7	1.0
avg. length per sentence	238.9	234.0	199.8	170.1	140.2	104.4	84.3	61.9
extension	14.6	15.2	12.6	10.5	8.4	5.9	4.5	3.0
avg. number of extensions per sentence	16.3	15.4	15.8	16.2	16.7	17.8	18.9	20.7

user–system interactions is reduced compared to the baseline.

5.4 Effect of Using Prefix Information

We ran three different sets of experiments to study the effect of the methods using information from a given prefix:

- restricting the calculation of confidence to the untranslated source words as introduced in equation 2,
- adapting the confidence threshold,
- the combination of both.

This was compared to an interactive system applying confidence estimation without exploiting the prefix information.

Table 4 contains an assessment of prediction quality in terms of precision and recall for the modified confidence estimation (labeled as "source") and for the adaptation of the confidence threshold (labeled as "threshold"). When only adapting the confidence threshold, the recall of the proposed extension increases, especially for higher threshold

Table 4. Effect of using prefix information for confidence threshold or source sentence on Recall [%] and Precision [%] for different values of the confidence threshold on the French→English test set. The confidence scaling factor is set to 0.8 (optimal w.r.t. F_{pred}).

confidence threshold			0.0	0.1	0.3	0.5	0.7	1.0
Recall [%]	use prefix:	no	64.9	64.2	62.9	61.3	59.8	57.8
		threshold	64.9	64.2	63.0	61.7	60.6	59.1
		source	64.9	64.2	62.9	61.4	60.1	58.2
Precision [%]	use prefix:	no	64.3	66.2	72.8	78.3	80.9	82.7
		threshold	64.3	66.1	71.7	76.2	79.0	80.9
		source	64.3	67.3	74.1	79.1	81.3	82.9

values. This is due to the fact that fewer correct words get discarded. But on the other hand, precision decreases significantly.

Confidence estimation based on a restricted source sentence as described in equation 2 in section 4.3 yields a slight degradation of recall, but a significant improvement in terms of precision, especially for low threshold values. The reason for this is that translations of source words which have been covered already by the prefix will now be correctly discarded.

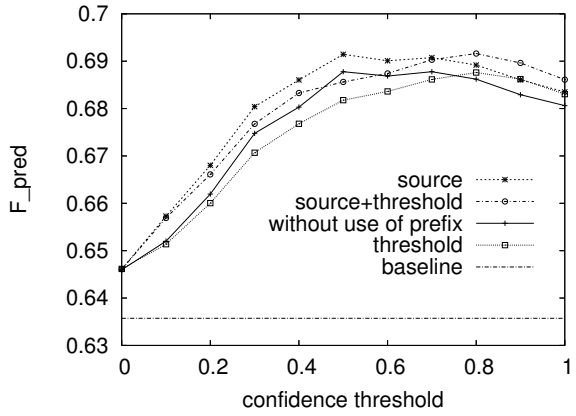

Figure 2. Effect of using prefix information for confidence threshold and source sentence on F_{pred} for different threshold values on the French→English test set. The confidence scaling factor is set to 0.8 (optimal w.r.t. F_{pred}).

The effect of restricting the source sentence to a prefix, adapting the threshold and their combination on F_{pred} is shown in figure 2. We see that the the calculation of word confidences according to a restricted source sentence ("source") consistently improves the predictions: The F-measure F_{pred} increases over the system without use of prefix information for all threshold values. Additionally, the adaptation of the confidence threshold can compensate for the negative effect of setting the threshold too high: For threshold values above 0.7, the system with threshold adaptation performs better than that without.

The combination of those two methods achieves the best results in terms of prediction F-measure for threshold values of 0.7 and higher, whereas the restriction of the source sentence performs best for lower threshold values.

An example of system output with and without consideration of the given prefix is given in table 5. The system that does not take the prefix into account proposes the word "System" as next extension, although the source word "système" has been translated already. The reason for this is that it has a high confidence w.r.t. the complete input sentence whereas the correct target word "Setup" has not. The improved system that determines the confidence only over the uncovered source words "Réglage du" is able to predict the correct extension "Setup".

5.5 Results on English–German

In order to verify the gain in the interactive system's performance, we ran additional experiments on a second language pair which is different from French→English in terms of strucure and complexity. We chose English→German which was also investigated in TransType2, for corpus statistics see table 1. We tested the system setup which proved best in the French→English experiments, i.e.

- selection of words in search step 3 (cf. section 4)

Table 5. Effect of the calculating confidence over a restricted source sentence on the proposed extensions. Example simulated interactive mode. The confidence scaling factor is set to 0.8, and the confidence threshold is 0.5 (both optimal w.r.t. F_{pred}). Example taken from the French→English test set.

source	Reportez-vous au chapitre 9 - Réglage du système
reference	Refer to Chapter 9 - System Setup
prefix	Refer to Chapter 9 - System S
system without use of prefix	ystem
system with use of prefix	etup

- rejection words in all search steps
- using the knowledge contained in the correct prefix for adaptation of the confidence estimation (see equation 2 in section 4.3)

The performance of this system on the English→German test set terms of F_{pred} is shown in figure 3. We see taht the gain in terms of F_{pred} is even higher than for French→English: The system improves by 13.4% absolute over the baseline.

When analyzing the number of user–system interactions presented in figure 4, one sees a significant reduction over the baseline for almost all threshold values.

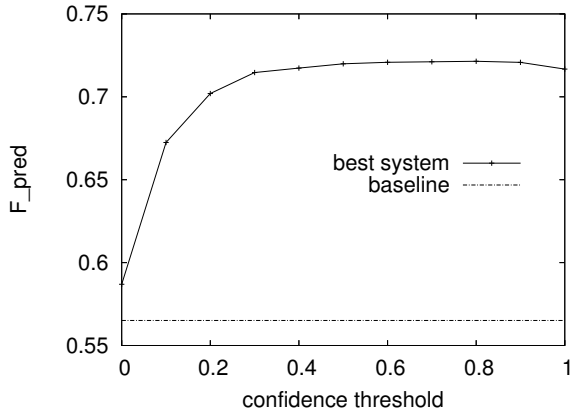

Figure 3. F_{pred} of the best system compared to the baseline system on the English→German test set. The confidence scaling factor is set to 0.8, and the threshold varies from 0 to 1.

6 Conclusion

We presented different novel ways of applying word-level confidence measures in an interactive statistical machine translation system. The system

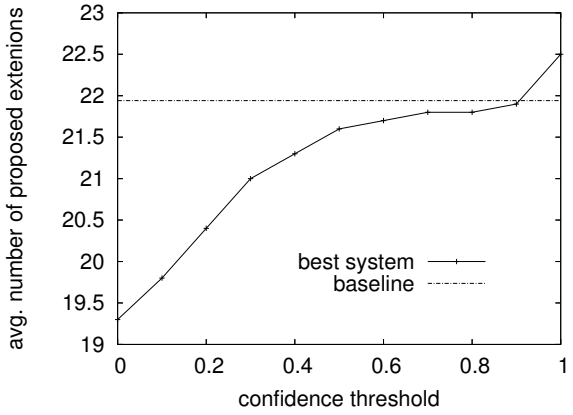

Figure 4. Average number of extensions proposed by the best system compared to the baseline system on the English→German test set. The confidence scaling factor is set to 0.8, and the threshold varies from 0 to 1.

is a state-of-the-art SMT system that suggests translations and adjusts them to a prefix that has already been typed by a user. We showed that the correctness of the translations predicted in this process can be significantly improved through rejection and selection of words based on their confidence.

Additionally, the confidence estimation has been modified such that it takes the given prefix into account. This resulted in an improvement of the system's performance as well. The results were evaluated using a metric measuring prediction accuracy and the keystrokes that a human user saves in typing the translation. The gain in prediction performance is 5.6% absolute over the baseline on the French→English test corpus, and 13.4% absolute on the English→German test set.

Future research will aim at the improvement of the confidence measure integrated into the interactive system as well as refined ways of exploitation of the given prefix.

Acknowledgement

This work was partly funded by the European Union under the RTD project TransType2 (IST–2001–32091, http://tt2.atosorigin.es), and by the Deutsche Forschungsgemeinschaft (DFG) under the project "Statistical Methods for Written Language Translation" (Ne572/5).

7 References

Atos Origin Spain, Intituto Tecnológico de Informática, Rheinisch Westfälische Technische Hochschule Aachen - Lehrstuhl für Informatik VI, Recherche Appliquée en Linguistique Informatique Laboratory - University of Montreal, Celer Soluciones, Société Gamma, and Xerox Research Centre Europe. 2002. TransType2 - computer assisted translation. `http://tt2.atosorigin.es/`.

O. Bender, R. Zens, E. Matusov, and H. Ney. 2004. Alignment Templates: the RWTH SMT System. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pp. 79–84, Kyoto, Japan, September.

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 *COLING '04: The 20th Int. Conf. on Computational Linguistics*, pp. 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.

J. Civera, J.M. Vilar, E. Cubel, A.L. Lagarda, S. Barrachina, E. Vidal, F. Casacuberta, D. Picó, and J. González. 2004. From machine translation to computer assisted translation using finite-state models. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Barcelona, Spain, July.

G. Foster, P. Isabelle, and P. Plamondon. 1996. Word completion: A first step toward target-text mediated IMT. In *COLING '96: The 16th Int. Conf. on Computational Linguistics*, pp. 394–399, Copenhagen, Denmark, August.

G. Foster, P. Isabelle, and P. Plamondon. 1997. Target-text mediated interactive machine translation. *Machine Translation*, 12(1):175–194.

S. Gandrabur and G. Foster. 2003. Confidence estimation for text prediction. In *Proc. Conf. on Natural Language Learning (CoNLL)*, pp. 95–102, Edmonton, Canada, May.

P. Langlais, G. Foster, and G. Lapalme. 2000. TransType: a computer-aided translation typing system. In *Workshop on Embedded Machine Translation Systems*, pp. 46–51, Seattle, Wash., May.

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

F.J. Och, R. Zens, and H. Ney. 2003. Efficient search for interactive statistical machine translation. In *Proc. 10th Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL)*, pp. 387–393, Budapest, Hungary, April.

C. Quirk. 2004. Training a sentence-level machine translation confidence metric. In *Proc. of the Fourth Int. Conf. on Language Resources and Evaluation (LREC)*, pp. 825–828, Lisbon, Portugal, May.

N. Ueffing and H. Ney. 2004. Bayes decision rules and confidence measures for statistical machine translation. In *EsTAL - España for Natural Language Processing*, pp. 70–81, Alicante, Spain, October. Lecture Notes in Computer Science, Springer Verlag.

N. Ueffing, F.J. Och, and H. Ney. 2002. Generation of word graphs in statistical machine translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pp. 156–163, Philadelphia, PA, July.