On the Use of Confidence Measures within an Interactive-predictive Machine Translation System

Jesús González-Rubio
Inst. Tec. de Informática
Univ. Politéc. de Valencia
46021 Valencia, Spain
jegonzalez@iti.upv.es

Daniel Ortiz-Martínez
Dpto. de Sist. Inf. y Comp.
Univ. Politéc. de Valencia
46021 Valencia, Spain
dortiz@dsic.upv.es

Francisco Casacuberta
Dpto. de Sist. Inf. y Comp.
Univ. Politéc. de Valencia
46021 Valencia, Spain
fcn@dsic.upv.es

Abstract

In this work, we address the question of how to integrate confidence measures into a interactive-predictive machine translation system and reduce user effort. Specifically, we propose to use word confidence measures to aid the user in validating correct prefixes from the outputs given by the system. Experimental results obtained on a corpus of the Bulletin of the European Union show that confidence information can help to reduce user effort.

1 Introduction

The research in the field of *machine translation* (MT) aims to develop computer systems which are able to translate text or speech without human intervention. However, present translation technology has not been able to deliver fully automated high-quality translations (Kay, 1997; Hutchins, 1999; Arnold, 2003). Typical solutions to improve the quality of the translations supplied by an MT system require manual post-editing. This serial process prevents the MT system from taking advantage of the knowledge of the human translator and the human translator can not take advantage of the adapting ability of the MT system.

An alternative way to take advantage of the existing MT technologies is to use them in *collaboration* with human translators within a *computer-assisted translation* (CAT) or *interactive* framework (Isabelle and Church, 1997). Interactivity in CAT has been explored for a long time. Systems have been designed to interact with human translators in order to solve ambiguities or update user dictionaries (Slocum, 1985; Whitelock et al., 1986).

An important contribution to CAT technology was pioneered by the *TransType* project (Foster et al., 1997; Langlais and Lapalme, 2002; Foster et al., 2002). It entailed a focus shift in which interaction directly aimed at the production of the target text, rather than at the disambiguation of the source text, as in former interactive systems. The idea proposed in that work was to embed data driven MT techniques within the interactive translation environment. Following the TransType ideas, Barrachina et al. (2009) proposed, in the *TransType-2* project, the use of fully-fledged statistical MT (SMT) systems to produce full target sentences hypotheses, or portions thereof, which can be accepted or amended by a human translator. Each correct text segment is then used by the MT system as additional information to achieve improved suggestions. More specifically, in each iteration, a prefix[1] of the target sentence is fixed by the human translator and, in the next iteration, the system predicts a best (or N-best) translation suffix(es)[1] to complete this prefix. This process is known as *Interactive-predictive Machine Translation* (IMT). In this paper, we also focus on the IMT approach to CAT.

Figure 1 illustrates a typical IMT session. Initially, the user is given an input sentence $\mathbf{f}$ to be translated. The provided reference $\mathbf{e}$ is the translation that the user would like to achieve at the end of the IMT session. At iteration 0, the user does not supply any correct text prefix to the system, for this reason the prefix $\mathbf{e}_p$ is shown as empty. Therefore,

[1]The terms prefix and suffix denote any substring at the beginning and end (respectively) of a string of characters, with no implication of morphological significance as is usually implied by these terms in linguistics.

SOURCE (f):		Para encencer la impresora:
REFERENCE (e):		To power on the printer:
ITER-0	$\mathbf{e}_p$	( )
	$\hat{\mathbf{e}}_s$	*To switch on a printer:*
ITER-1	$\mathbf{a}$	To
	$\mathbf{k}$	**power**
	$\mathbf{e}_p$	To power
	$\hat{\mathbf{e}}_s$	*on a printer:*
ITER-2	$\mathbf{a}$	on
	$\mathbf{k}$	**the**
	$\mathbf{e}_p$	To power on the
	$\hat{\mathbf{e}}_s$	*printer:*
FINAL	$\mathbf{a}$	printer:
	$\mathbf{k}$	#
	$\mathbf{e}_p = \mathbf{e}$	To power on the printer:

Figure 1: IMT session to translate a Spanish sentence into English. System suggestions are in italics, accepted prefixes are printed in normal font and user inputs are in boldface font.

the IMT system has to provide an initial complete translation $\hat{\mathbf{e}}_s$, as if it were a conventional SMT system. In the next iteration, the user accepts a preffix of this suffix $\mathbf{a}$ and introduces a correction $\mathbf{k}$. This being done, the system suggests a new suffix hypothesis $\hat{\mathbf{e}}_s$, subject to $\mathbf{e}_p \equiv \mathbf{ak}$. Again, the user validates a new prefix, introduces a new correction and so forth. The process continues until the whole sentence is correct. A correct sentence is validated by introducing the special word "#".

As the reader could devise from the IMT session described above, IMT aims at reducing the effort and increasing the productivity of translators, while preserving high-quality translation.

In this work, we intend to further reduce the user effort. As explained above, in each iteration, the user is asked to validate a prefix of the hypothesis generated by the system and then, to make a correction. To do that, the user only has information about the source sentence to be translated. We propose to provide the user with information about the correctness for each word in the suffix. This *confidence measure* (CM) will guide the user to locate possible translation errors in the sufixes given by the IMT system.

2 Confidence Measures

Sentences generated by a MT system are often incorrect but may contain correct substrings. Using CMs allow to identify these correct substrings and find possible errors. For this purpose, each word in the generated target sentence is assigned a value expressing the confidence that it is correct. Confidence estimation can be seen as a conventional pattern classification problem in which a feature vector is obtained for each hypothesised word in order to classify it as either correct or incorrect. Confidence estimation have been extensively studied for speech recognition. Only recently have researchers started to investigate CMs for MT (Gandrabur and Foster, 2003; Blatz et al., 2004; Quirk, 2004; Ueffing and Ney, 2007; Sanchis et al., 2007; Specia et al., 2009).

Different TransType-style MT systems use confidence information to improve translation prediction accuracy (Foster et al., 2002; Gandrabur and Foster, 2003; Ueffing and Ney, 2005). In this work, we propose a focus shift in which confidence information is used to aid the user in validating correct prefixes by locating incorrectly translated words in the sufixes given by the IMT system.

2.1 Selecting a Confidence Measure for IMT

Two problems have to be solved in order to compute CMs. First, suitable confidence features have to be computed. Second, a binary classifier has to be defined, which decides whether a word is correct or not.

In this work, we implement a word CM based on the IBM Model 1 (Brown et al., 1993), similar to the one described in (Blatz et al., 2004). We choose this because it relies only on the source

SOURCE (f):		Para encencer la impresora:
REFERENCE (e):		To power on the printer:
ITER-0	$\mathbf{e}_p$	( )
	$\hat{\mathbf{e}}_s$	*To switch on a printer:*
ITER-1	$\mathbf{a}$	To switch on
	$\mathbf{k}$	**the**
	$\mathbf{e}_p$	To switch on the
	$\hat{\mathbf{e}}_s$	*printer:*
FINAL	$\mathbf{a}$	printer:
	$\mathbf{k}$	#
	$\mathbf{e}_p \equiv \mathbf{e}$	To `switch` on the printer:

Figure 2: IMT session with confidence information using our proposed user simulation. System suggestions are in italics, accepted prefixes are printed in normal font and user inputs are in boldface font. Words classified as incorrect are displayed underlined and translation errors are printed in typewriter font. The final output is different from the reference translation $\mathbf{e}$, but it is also a correct translation of the source sentence $\mathbf{f}$.

sentence and the proposed extension, and not on an N-best list or an additional confidence estimation layer as many other word CMs do. Thus, it can be calculated very fast during search, which is crucial given the time constraints of the IMT systems. Moreover, its performance in identifying correct words is similar to that of other word CMs as the results presented in (Blatz et al., 2003; Blatz et al., 2004; Sanchis et al., 2007) show. However, we modified this CM by replacing the *average* by the *maximal* lexicon probability, because work by Ueffing and Ney (2005) show that the average is dominated by this maximum. The confidence value of word e_i, $c(e_i)$, is then given by

$$c(e_i) = \max_{0 \leq j \leq J} p(e_i|f_j) \;, \qquad (1)$$

where $p(e_i|f_j)$ is the lexicon probability based on the IBM Model 1, f_0 is the empty source word and J is the number of words in the source sentence. Ueffing and Ney (2005) report that even this relatively simple CM yields a significant improvement in the quality of the suffixes proposed by an IMT system.

After computing the confidence value, each word is classified as either correct or incorrect, depending on whether its confidence exceeds or not a clasiffication threshold.

3 IMT with Confidence Measures

In the IMT approach (see Figure 1), the user interaction with the IMT system consists on validating a correct prefix for each suffix $\hat{\mathbf{e}}_s$ given by the system. To do that, the user has to check the correctness of each word in the given suffix looking for the first incorrectly translated word. We propose the use of CMs as a new source of information to aid the user in locating these incorrectly translated words.

In a conventional IMT system, the only information available to the user is the source sentence to be translated, so, all the words of the target sentence are equally likely to be correct or incorrect. In contrast, we propose to provide the user with information about the correctness of each of the words in the suffix. In our proposal, the user has more available information which can help her to easily validate the correct prefix.

To appropriately evaluate the impact of providing the user with confidence information within the IMT scenario, experimentation involving human translators should be carried out. Unfortunately, such a user study would be very costly. Because of this, we are forced to carry out experimentation simulating the human translators. This user simulation does not intend to exactly imitate the behaviour of real IMT users, but to test if confidence information may be useful for a human translator within the IMT process. Anyway, experimentation involving human translators will be carried out in the future.

3.1 User Simulation

We want to study the impact of using CMs within the IMT process. To do that, we simulate a hu-

man translator that absolutely rely on the confidence information to validate correct prefixes from the suffixes given by the IMT system. To simulate such a human translator, we make two assumptions. First, we assume that the CM makes no mistakes in clasiffying words. Second, we assume that the user is always able to correct a word without taking into account the context of this word.

The first assumption implies that the user checks the correctness of only those words that are classified as incorrect, skipping the words classified as correct. Confidence estimation is not perfect, therefore some of the words may be misclassified, as a result, the output generated by our user simulation is not guaranteed to be equal to the reference.

The second assumption is a consequence of the first one. If we skip words that may be incorrect, the user should be capable of correcting each incorrect word even when the context of this word may be erroneous. We use the reference sentence to correct the words classified as incorrect, i.e. if the second word of a suffix needs to be corrected, we correct it with the word in the same position in the corresponding reference sentence.

We are aware that the above described assumptions may seem unrealistic, but they are made to simplify the IMT scenario in which the impact of using confidence information is to be evaluated.

Our user simulation is exemplified in Figure 2. At iteration 0, the system has classified the word *a* as incorrect (words classified as incorrect are displayed underlined in the example). With this information the user focuses her attention directly on the word *a* and corrects it, skipping the words "*To switch on*" that the system considers to be correct. Word *switch* is different from the reference word *power*, so, in this scenario, the final translation error will be greater than zero. At the second iteration there are no words classified as erroneous, so the user accepts the suffix without checking any of the suffix words. Following the conventional IMT approach, the user has to check the correctness of 5 words and correct two of them to obtain the desired translation, while in our simulation, the user has to check the correctness of only one word and correct it to obtain the final translation. In spite of the fact that this final translation is different from the one the user has in mind, it is a correct translation of the source sentence.

It is worth of notice that, in our user simulation, varying the value of the classification threshold allows to range from a fully automatic SMT approach (threshold equal to 0.0, all words are classified as correct) to a conventional IMT approach (threshold equal to 1.0, all words are classified as incorrect). The classification threshold value allows us to control the ratio between the user effort required by the IMT system and the expected final translation error, according to the requirements of the given translation task. For any threshold value lower than 1.0 our user simulation does not guarantee error free translations.

4 Experimentation

The aim of this experimentation was to study the impact of providing the user of an IMT system with confidence information. All the experiments were carried out using the user simulation described in section 3.1.

4.1 System evaluation

Automatic evaluation of results is a difficult problem in MT. In fact, it has evolved to a research field with its own identity. This is due to the fact that, given an input sentence, a great number of correct and different output sentences may exist. Hence, there is no sentence which can be considered ground truth, as it is the case in speech or text recognition. By extension, this problem is also applicable to our user simulation. Moreover, we additionally have to deal with the problem of measuring the user effort.

In this paper, we report our results as measured by *Word Stroke Ratio* (WSR) (Tomás and Casacuberta, 2006). WSR is used in the context of IMT to measure the effort required by the user to generate her translations. WSR is computed as the quotient between the number of word-strokes a user would need to perform in order to achieve the translation she has in mind and the total number of words in the sentence. In this context, a word-stroke is interpreted as a single action, in which the user types a complete word, and is assumed to have constant cost. Moreover, each word-stroke also takes into account the cost incurred by the user when reading the new suffix provided by the system.

In addition, and because our user simulation allows differences between its output and the reference translation, we will also present translation quality results in terms of *Translation Edit Rate* (TER) (Snover et al., 2006) and *BiLingual Evaluation Understudy* (BLEU) (Papineni et al., 2002).

		Spanish	English
Train	Sentences	214.5K	
	Running words	5.8M	5.2M
	Vocabulary	97.4K	83.7K
Dev.	Sentences	400	
	Running words	11.5K	10.1K
	Perplexity (trigrams)	46.1	59.4
Test	Sentences	800	
	Running words	22.6K	19.9K
	Perplexity (trigrams)	45.2	60.8

Table 1: Statistics of the Spanish–English EU corpora. K and M denote thousands and millions of elements respectively.

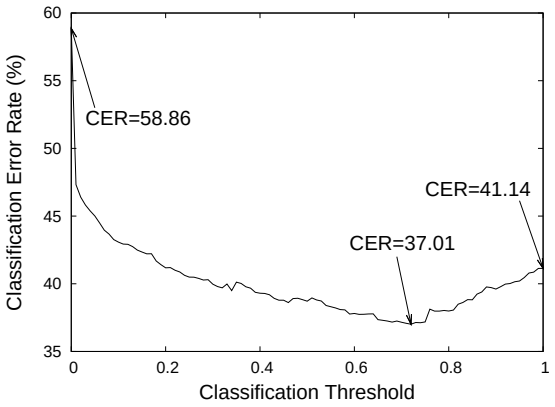

Figure 3: CER for different classification threshold values when translating from Spanish into English.

TER is calculated as the number of edit operations (insertions, deletions and substitutions of single words and shifts of word sequences) to convert the system translation into the reference translation. BLEU computes a geometric mean of the precision of n-grams multiplied by a factor to penalise short sentences.

Finally, to evaluate the performance of the selected CM we use the *Classification Error Rate* (CER). This metric is defined as the number of classification errors divided by the total number of classified words.

4.2 Experimental Setup

Our experiments were carried out on the EU corpora (Barrachina et al., 2009). The EU corpora were extracted from the Bulletin of the European Union, which is publicly available on the Internet. The EU corpora are composed of sentences given in three different language pairs. Here, we will focus on the Spanish–English part of the EU corpora. The corpus is divided into three separate sets: one for training, one for development, and one for test. The figures of the corpus can be seen in Table 1.

As a first step, be built a SMT system to translate from Spanish into English. This was done by means of the Thot toolkit (Ortiz et al., 2005), which is a complete system for building phrase-based SMT models. This toolkit involves the estimation from the training set of different statistical models, which are combined in a log-linear fashion by adjusting a weight for each of them by means of the MERT (Och, 2003) procedure, optimising the BLEU score on the development partition.

The IMT system which we have implemented relies on the use of word graphs (Ueffing et al., 2002) to efficiently compute the suffix for a given prefix. A word graph has to be generated for each sentence to be interactively translated. For this purpose, we used a multi-stack phrase-based decoder which will be distributed in the near future together with the Thot toolkit. We discarded the use of the state-of-the-art Moses toolkit (Koehn et al., 2007) because preliminary experiments performed with it revealed that the decoder by Ortiz-Martínez et al. (2005) performs clearly better when used to generate word graphs for their use in IMT. In addition, we performed an experimental comparison in regular SMT, and found that the performance difference was negligible. The decoder was set to only consider monotonic translation, since in real IMT scenarios considering non-monotonic translation leads to excessive response time for the user.

Finally, the obtained word graphs were used in our user simulation to produce the translations of the sentences in the test set, measuring WSR, TER and BLEU.

4.3 Word Confidence Classification Results

We carried out an experimentation intended to study the performance of the CM in classifying the words as correct or incorrect. In order to evaluate the classification performance of the CM, a corpus is needed where each word is tagged as correct or incorrect. We carried out a conventional IMT session to produce the reference translations and use the user interactions with the system to tag the words as correct or incorrect. For example, in the IMT session in Figure 1, at iteration 1 word *To* is tagged as correct because the user marked it as a valid prefix and word *switch* is tagged as in-

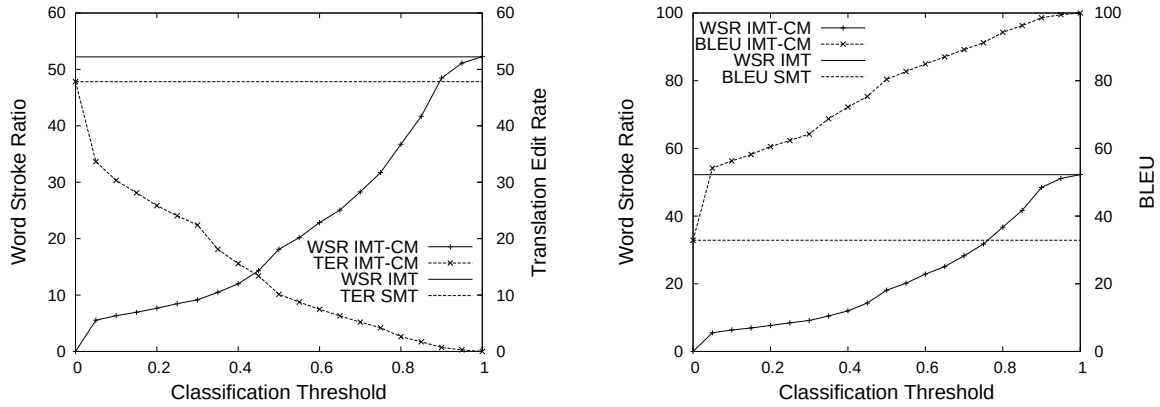

Figure 4: TER (left) and BLEU (right) translation scores against WSR for different values of the confidence classification threshold when translating from Spanish into English.

correct because the user corrects it with the word *power*. At iteration 2, word *on* is tagged as correct and word *a* as incorrect. Finally, word *printer:* is tagged as correct. Once the words are tagged, confidence classification is performed for a certain classification threshold and the CER score for this threshold is calculated.

Figure 3 displays CER for different values of the classification threshold. The two extreme values 0.0 and 1.0 imply that the CM does not add information about the correctness of the words in the suffix. Specifically, a threshold value equal to 0.0 classifies all the target words as correct, whereas a threshold value equal to 1.0 classifies all the target words as incorrect.

According to Figure 3, best CER score was obtained for a threshold value of 0.75. This threshold value allows to achieve better CER score than that obtained using a threshold value of 1.0. Since a threshold value of 1.0 corresponds to the conventional IMT system, we conclude that providing the user with confidence information is better than not providing confidence information at all.

4.4 User Simulation IMT Results

In the previous section, we have seen that confidence information is useful to detect incorrectly translated words, and so, may make the user interaction with the IMT system easier. One advantage of integrating CMs within an IMT system is their ability to achieve a trade-off between the required user effort and the expected final translation error.

In this section, we present a series of experiments ranging the value of the classification threshold between 0.0 (unsupervised SMT system) and 1.0 (conventional IMT system). For each threshold value, we calculated the effort of our simulated user in terms of WSR, and the translation quality of the final output as measured by TER and BLEU.

Figure 4 shows WSR (WSR IMT-CM), TER (TER IMT-CM) and BLEU (BLEU IMT-CM) scores obtained by our user simulation for different classification threshold values. Additionally, we also show the TER and BLEU scores (TER SMT and BLEU SMT) obtained by a fully automatic SMT system as translation quality baselines, and the WSR score (WSR IMT) obtained by a conventional IMT system as user effort baseline.

Figure 4 shows a smooth transition between the unsupervised SMT system and the conventional IMT system. As we raised the threshold value, more words were marked as incorrect, and therefore, more words were suitable for correction. According to Figure 4, using the best threshold value (0.75) in Figure 3, we can achieve a translation error as low as 4 TER points by correcting only 30% absolute of the words. This constitutes a WSR reduction of 40% relative with respect to the standard IMT approach and a BLEU improvement of almost 60 points with respect to the unsupervised SMT system.

It is worth of notice that the experimentation is carried out simulating a user whose decisions are absolutely guided by the confidence information. The user effort savings and the improvements over the SMT translation quality displayed in Figure 4, confirm that confidence information can aid a human translator in making her decisions within the IMT process.

5 Concluding Remarks

In this work, we proposed to enrich the IMT framework with confidence information. Since an experimentation involving human user would be very costly, we were forced to design a simulation of the human users to test our proposal. This user simulation was not intended to reproduce a real IMT user, but to test if confidence information may be useful for a real IMT user.

Experimentation results show that confidence information can aid real users to locate incorrectly translated words, making easier for them to validate correct prefixes within an IMT framework. According to our user simulation, a 40% reduction in the WSR was obtained with respect to the conventional IMT system. In addition, an improvement of 60 BLEU points is also achieved with respect to the SMT system.

As future work, we plan to perform a human evaluation to verify the results obtained with our user simulation.

Acknowledgements

Work supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018) and the FPU scholarship AP2006-00691. Also supported by the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project and by the Generalitat Valenciana under grant Prometeo/2009/014.

References

Arnold, Doug, 2003. *Computers and Translation: A translator's guide*, chapter 8, pages 119–142.

Barrachina, Sergio, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, and Enrique Vidal. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Blatz, Jonh, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence estimation for machine translation.

Blatz, Jonh, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the International Conference on Computational Linguistics*, page 315.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Foster, George, Pierre Isabelle, and Pierre Plamondon. 1997. Target-text mediated interactive machine translation. *Machine Translation*, 12:12–175.

Foster, George, Philippe Langlais, and Guy Lapalme. 2002. User-friendly text prediction for translators. In *Proceedings of the conference on Empirical methods in natural language processing*, pages 148–155.

Gandrabur, Simona and George Foster. 2003. Confidence estimation for text prediction. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 315–321.

Hutchins, Jonh. 1999. Retrospect and prospect in computer-based translation. In *Proceedings of the Machine Translation Summit*, pages 30–44.

Isabelle, Pierre and Ken Church. 1997. Special issue on new tools for human translators. *Machine Translation*, 12(1–2).

Kay, Martin. 1997. It's still the proper place. *Machine Translation*, 12(1-2):35–38.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Association for Computational Linguistics meeting*, pages 177–180.

Langlais, Philippe and Guy Lapalme. 2002. Transtype: Development-evaluation cycles to boost translator's productivity. *Machine Translation*, 15(4):77–98.

Och, Franz J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the Association for Computational Linguistics meeting*, pages 160–167.

Ortiz, Daniel, Ismael García-Varea, and Francisco Casacuberta. 2005. Thot: a toolkit to train phrase-based statistical translation models. In *Proceedings of the Machine Translation Summit*, pages 141–148.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of MT. In *Proceedings of the Association for Computational Linguistics meeting*, pages 311–318.

Quirk, Chris. 2004. Training a sentence-level machine translation confidence metric. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 825–828.

Sanchis, Alberto, Alfons Juan, and Enrique Vidal. 2007. Estimation of confidence measures for machine translation. In *Proceedings of the Machine Translation Summit*, pages 407–412.

Slocum, Jonathan. 1985. A survey of machine translation: Its history, current status, and future prospects. *Computational Linguistics*, 11(1):1–17.

Snover, Mattew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc of the Association for Machine Translation in the Americas meeting*, pages 223–231.

Specia, Lucia, Marco Turchi, Zhuoran Wang, John Shawe-Taylor, and Craig Saunders. 2009. Improving the confidence of machine translation quality estimates. In *Proceedings of the Machine Translation Summit*.

Tomás, Jesús and Francisco Casacuberta. 2006. Statistical phrase-based models for interactive computer-assisted translation. In *Proceedings of the International Conference on Computational Linguistics*, pages 835–841.

Ueffing, Nicola and Hermann Ney. 2005. Application of word-level confidence measures in interactive statistical machine translation. In *Proceedings of the European Association for Machine Translation conference*, pages 262–270.

Ueffing, Nicola and Hermann Ney. 2007. Word-level confidence estimation for machine translation. *Computational Linguistics*, 33(1):9–40.

Ueffing, Nicola, Franz J. Och, and Hermann Ney. 2002. Generation of word graphs in statistical machine translation. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pages 156–163.

Whitelock, P.J., M. Wood, B.J. Chandler, N. Holden, and H.J. Horsfall. 1986. Strategies for interactive machine translation: the experience and implications of the umist japanese project. In *Proceedings of the Association for Computational Linguistics*, pages 329–334.