Balancing User Effort and Translation Error in Interactive Machine Translation Via Confidence Measures

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

This work deals with the application of confidence measures within an interactive-predictive machine translation system in order to reduce human effort. If a small loss in translation quality can be tolerated for the sake of efficiency, user effort can be saved by interactively translating only those initial translations which the confidence measure classifies as incorrect. We apply confidence estimation as a way to achieve a balance between user effort savings and final translation error. Empirical results show that our proposal allows to obtain almost perfect translations while significantly reducing user effort.

1 Introduction

In *Statistical Machine Translation* (SMT), the translation is modelled as a decission process. For a given source string $f_1^J = f_1 \dots f_j \dots f_J$, we seek for the target string $e_1^I = e_1 \dots e_i \dots e_I$ which maximises posterior probability:

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

Within the *Interactive-predictive Machine Translation* (IMT) framework, a state-of-the-art SMT system is employed in the following way: For a given source sentence, the SMT system fully automatically generates an initial 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 $e_1^i = e_1 \dots e_i$ into account. These steps are repeated until the whole input sentence has been correctly translated. In the resulting decision rule, we maximise over all possible extensions e_{i+1}^I of e_1^i:

$$\hat{e}_{i+1}^{\hat{I}} = \underset{I, e_{i+1}^I}{\operatorname{argmax}} Pr(e_{i+1}^I | e_1^i, f_1^J) \; . \qquad (2)$$

An implementation of the IMT famework was performed in the TransType project (Foster et al., 1997; Langlais et al., 2002) and further improved within the TransType2 project (Esteban et al., 2004; Barrachina et al., 2009).

IMT aims at reducing the effort and increasing the productivity of translators, while preserving high-quality translation. In this work, we integrate *Confidence Measures* (CMs) within the IMT framework to further reduce the user effort. As will be shown, our proposal allows to balance the ratio between user effort and final translation error.

1.1 Confidence Measures

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; Ueffing and Ney, 2007).

Different TransType-style MT systems use confidence information to improve translation prediction accuracy (Gandrabur and Foster, 2003; Ueffing and Ney, 2005). In this work, we propose a focus shift in which CMs are used to modify the interaction between the user and the system instead of modify the IMT translation predictions.

To compute CMs we have to select suitable confidence features and define a binary classifier. Typically, the classification is carried out depending on whether the confidence value exceeds a given threshold or not.

2 IMT with Sentence CMs

In the conventional IMT scenario a human translator and a SMT system collaborate in order to obtain the translation the user has in mind. Once the user has interactively translated the source sentences, the output translations are error-free. We propose an alternative scenario where not all the source sentences are interactively translated by the user. Specifically, only those source sentences

Proceedings of the ACL 2010 Conference Short Papers, pages 173–177,
Uppsala, Sweden, 11-16 July 2010.

whose initial fully automatic translation are incorrect, according to some quality criterion, are interactively translated. We propose to use CMs as the quality criterion to classify those initial translations.

Our approach implies a modification of the user-machine interaction protocol. For a given source sentence, the SMT system generates an initial translation. Then, if the CM classifies this translation as correct, we output it as our final translation. On the contrary, if the initial translation is classified as incorrect, we perform a conventional IMT procedure, validating correct prefixes and generating new suffixes, until the sentence that the user has in mind is reached.

In our scenario, we allow the final translations to be different from the ones the user has in mind. This implies that the output may contain errors. If a small loss in translation can be tolerated for the sake of efficiency, user effort can be saved by interactively translating only those sentences that the CMs classify as incorrect.

It is worth of notice that our proposal can be seen as a generalisation of the conventional IMT approach. Varying the value of the CM classification threshold, we can range from a fully automatic SMT system where all sentences are classified as correct to a conventional IMT system where all sentences are classified as incorrect.

2.1 Selecting a CM for IMT

We compute sentence CMs by combining the scores given by a word CM based on the IBM model 1 (Brown et al., 1993), similar to the one described in (Blatz et al., 2004). We modified this word CM by replacing the *average* by the *maximal* lexicon probability, because the average is dominated by this maximum (Ueffing and Ney, 2005). We choose this word CM because it can be calculated very fast during search, which is crucial given the time constraints of the IMT systems. Moreover, its performance is similar to that of other word CMs as results presented in (Blatz et al., 2003; Blatz et al., 2004) show. The word confidence value of word e_i, $c_w(e_i)$, is given by

$$c_w(e_i) = \max_{0 \leq j \leq J} p(e_i|f_j) \ , \qquad (3)$$

where $p(e_i|f_j)$ is the IBM model 1 lexicon probability, and f_0 is the empty source word.

From this word CM, we compute two sentence CMs which differ in the way the word confidence scores $c_w(e_i)$ are combined:

		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.

MEAN CM ($c_M(e_1^I)$) is computed as the geometric mean of the confidence scores of the words in the sentence:

$$c_M(e_1^I) = \sqrt[I]{\prod_{i=1}^{I} c_w(e_i)} \ . \qquad (4)$$

RATIO CM ($c_R(e_1^I)$) is computed as the percentage of words classified as correct in the sentence. A word is classified as correct if its confidence exceeds a word classification threshold τ_w.

$$c_R(e_1^I) = \frac{|\{e_i \ / \ c_w(e_i) > \tau_w\}|}{I} \qquad (5)$$

After computing the confidence value, each sentence is classified as either correct or incorrect, depending on whether its confidence value exceeds or not a sentence clasiffication threshold τ_s. If $\tau_s = 0.0$ then all the sentences will be classified as correct whereas if $\tau_s = 1.0$ all the sentences will be classified as incorrect.

3 Experimentation

The aim of the experimentation was to study the possibly trade-off between saved user effort and translation error obtained when using sentence CMs within the IMT framework.

3.1 System evaluation

In this paper, we report our results as measured by *Word Stroke Ratio* (WSR) (Barrachina et al., 2009). WSR is used in the context of IMT to measure the effort required by the user to generate her

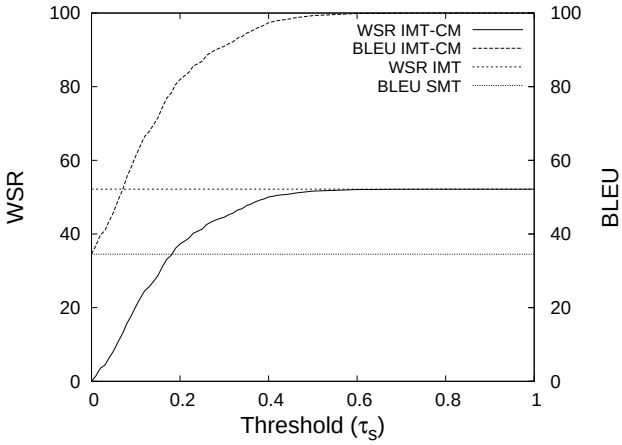

Figure 1: BLEU translation scores versus WSR for different values of the sentence classification threshold using the MEAN CM.

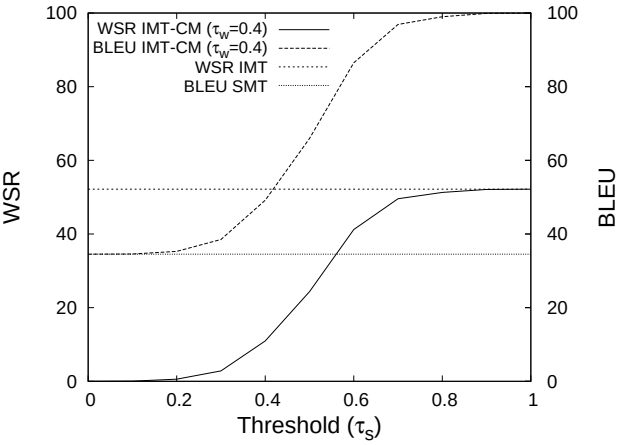

Figure 2: BLEU translation scores versus WSR for different values of the sentence classification threshold using the RATIO CM with $\tau_w = 0.4$.

translations. WSR is computed as the ratio between the number of word-strokes a user would need 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.

Additionally, and because our proposal allows differences between its output and the reference translation, we will also present translation quality results in terms of *BiLingual Evaluation Understudy* (BLEU) (Papineni et al., 2002). BLEU computes a geometric mean of the precision of n-grams multiplied by a factor to penalise short sentences.

3.2 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. The EU corpora is 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 training, development and test sets. The main 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 in turn 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 set.

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 to use 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 better in terms of WSR when used to generate word graphs for their use in IMT (Sanchis-Trilles et al., 2008). Moreover, the performance difference in regular SMT is 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 within the IMT procedure to produce the reference translations in the test set, measuring WSR and BLEU.

3.3 Results

We carried out a series of experiments ranging the value of the sentence classification threshold τ_s, between 0.0 (equivalent to a fully automatic SMT system) and 1.0 (equivalent to a conventional IMT system), for both the MEAN and RATIO CMs. For each threshold value, we calculated the effort of the user in terms of WSR, and the translation quality of the final output as measured by BLEU.

src-1	DECLARACIÓN (No 17) relativa al derecho de acceso a la información
ref-1	DECLARATION (No 17) on the right of access to information
tra-1	DECLARATION (No 17) on the right of access to information
src-2	Conclusiones del Consejo sobre el comercio electrónico y los impuestos indirectos.
ref-2	Council conclusions on electronic commerce and indirect taxation.
tra-2	Council conclusions on e-commerce and indirect taxation.
src-3	participación de los países candidatos en los programas comunitarios.
ref-3	participation of the applicant countries in Community programmes.
tra-3	countries' involvement in Community programmes.

Example 1: Examples of initial fully automatically generated sentences classified as correct by the CMs.

Figure 1 shows WSR (WSR IMT-CM) and BLEU (BLEU IMT-CM) scores obtained varying τ_s for the MEAN CM. Additionally, we also show the BLEU score (BLEU SMT) obtained by a fully automatic SMT system as translation quality baseline, and the WSR score (WSR IMT) obtained by a conventional IMT system as user effort baseline. This figure shows a continuous transition between the fully automatic SMT system and the conventional IMT system. This transition occurs when ranging τ_s between 0.0 and 0.6. This is an undesired effect, since for almost a half of the possible values for τ_s there is no change in the behaviour of our proposed IMT system.

The RATIO CM confidence values depend on a word classification threshold τ_w. We have carried out experimentation ranging τ_w between 0.0 and 1.0 and found that this value can be used to solve the above mentioned undesired effect for the MEAN CM. Specifically, varying the value of τ_w we can stretch the interval in which the transition between the fully automatic SMT system and the conventional IMT system is produced, allowing us to obtain smother transitions. Figure 2 shows WSR and BLEU scores for different values of the sentence classification threshold τ_s using $\tau_w = 0.4$. We show results only for this value of τ_w due to paper space limitations and because $\tau_w = 0.4$ produced the smoothest transition. According to Figure 2, using a sentence classification threshold value of 0.6 we obtain a WSR reduction of 20% relative and an almost perfect translation quality of 87 BLEU points.

It is worth of notice that the final translations are compared with only one reference, therefore, the reported translation quality scores are clearly pessimistic. Better results are expected using a multi-reference corpus. Example 1 shows the source sentence (src), the reference translation (ref) and the final translation (tra) for three of the initial fully automatically generated translations that were classified as correct by our CMs, and thus, were not interactively translated by the user. The first translation (tra-1) is identical to the corresponding reference translation (ref-1). The second translation (tra-2) corresponds to a correct translation of the source sentence (src-2) that is different from the corresponding reference (ref-2). Finally, the third translation (tra-3) is an example of a slightly incorrect translation.

4 Concluding Remarks

In this paper, we have presented a novel proposal that introduces sentence CMs into an IMT system to reduce user effort. Our proposal entails a modification of the user-machine interaction protocol that allows to achieve a balance between the user effort and the final translation error.

We have carried out experimentation using two different sentence CMs. Varying the value of the sentence classification threshold, we can range from a fully automatic SMT system to a conventional IMT system. Empirical results show that our proposal allows to obtain almost perfect translations while significantly reducing user effort.

Future research aims at the investigation of improved CMs to be integrated in our IMT system.

Acknowledgments

Work supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018), the iTransDoc (TIN2006-15694-CO2-01) and iTrans2 (TIN2009-14511) projects 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

S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. Lagarda, H. Ney, J. Tomás, and E. Vidal. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2003. Confidence estimation for machine translation.

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kuesza, A. Sanchis, and N. Ueffing. 2004. Confidence estimation for machine translation. In *Proc. COLING*, page 315.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

J. Esteban, J. Lorenzo, A. Valderrábanos, and G. Lapalme. 2004. Transtype2: an innovative computer-assisted translation system. In *Proc. ACL*, page 1.

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

S. Gandrabur and G. Foster. 2003. Confidence estimation for text prediction. In *Proc. CoNLL*, pages 315–321.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL*, pages 177–180.

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

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

D. Ortiz, I. García-Varea, and F. Casacuberta. 2005. Thot: a toolkit to train phrase-based statistical translation models. In *Proc. MT Summit*, pages 141–148.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of MT. In *Proc. ACL*, pages 311–318.

G. Sanchis-Trilles, D. Ortiz-Martínez, J. Civera, F. Casacuberta, E. Vidal, and H. Hoang. 2008. Improving interactive machine translation via mouse actions. In *Proc. EMNLP*, pages 25–27.

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

N. Ueffing and H. Ney. 2007. Word-level confidence estimation for machine translation. *Comput. Linguist.*, 33(1):9–40.

N. Ueffing, F.J. Och, and H. Ney. 2002. Generation of word graphs in statistical machine translation. In *Proc. EMNLP*, pages 156–163.