New Confidence Measures for Statistical Machine Translation

Sylvain Raybaud, Caroline Lavecchia, David Langlois, Kamel Smaïli

NEW CONFIDENCE MEASURES FOR STATISTICAL MACHINE TRANSLATION

Sylvain Raybaud, Caroline Lavecchia, David Langlois, Kamel Smaïli
PAROLE team, LORIA, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy FRANCE
{*raybauds,lavecchi,langlois,smaili*}*@loria.fr*

Keywords: confidence measure, statistical machine translation, mutual information, linguistic features language model, n-gram language model

Abstract: A confidence measure is able to estimate the reliability of an hypothesis provided by a machine translation system. The problem of confidence measure can be seen as a process of testing : we want to decide whether the most probable sequence of words provided by the machine translation system is correct or not. In the following we describe several original word-level confidence measures for machine translation, based on mutual information, n-gram language model and lexical features language model. We evaluate how well they perform individually or together, and show that using a combination of confidence measures based on mutual information yields a classification error rate as low as 25.1% with an F-measure of 0.708.

1 INTRODUCTION

Statistical techniques have been widely used and remarkably successful in automatic speech recognition, machine translation and in natural language processing over the last two decades. This success is due to the fact that this approach is language independent and requires no prior knowledge, only large enough text corpora to estimate probability densities on. However statistical methods suffer from an intrinsic drawback: they only produce the result which is most likely given training and input data. It is easy to see that this will sometimes not be optimal with regard to human expectations. It is therefore important to be able to automatically evaluate the quality of the result: this can be handled by the different *confidence measures (CMs)* which have been proposed for machine translation.

In this paper we introduce new CMs to assess the reliability of translation results. The proposed CMs take advantage of the constituents of a translated sentence: n-grams, word triggers, and also word features.

1.1 A Brief Overview of Statistical Machine Translation Principle

In this framework the translation process is essentially the search for the most probable sentence in the target language given a sentence in the source language; let $\mathbf{f} = f_1,..,f_I$ be the source sentence (to be translated) and $\hat{\mathbf{e}} = e_1,..,e_J$ be the sentence generated by the system (target sentence):

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f}) \quad (1)$$

which is equivalent (using the Bayes rule) to:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} P(\mathbf{e})P(\mathbf{f}|\mathbf{e}) \quad (2)$$

In Equation 2, $P(\mathbf{e})$ is estimated from a *language model* and is supposed to estimate the correctness of the sentence ("is it a good sentence in the target language ?"), and $P(\mathbf{f}|\mathbf{e})$ is computed from a *translation model* and is supposed to reflect the accuracy of the translation ("does the generated sentence carry exactly the same information than the source sentence ?"). The language model is itself estimated on a large text corpus written in the target language, while the translation model is computed on a bilingual aligned

corpus (a text and its translation with line-wise correspondence). The decoder then generates the best hypothesis by making a compromise between these two probabilities.
Of course there are three main drawbacks to this approach: first the search space is so huge that exact computation of the optimum is intractable; second, even if it was, statistical models have inherent limitations which prevent them from being completely sound linguistically; finally, the probability distribution P can only be estimated on finite corpora, and therefore suffers from imprecision and data sparsity. Because of that, any SMT system sometimes produces erroneous translations. It is an important task to detect and possibly correct these mistakes, and this could be handled by confidence measures.

2 AN INTRODUCTION TO CONFIDENCE MEASURES

2.1 Motivation and Principle of Confidence Estimation

As said before, SMT systems make mistakes. A word's translation can be wrong, misplaced, or missing. Extra words can be inserted. A whole sentence can be wrong or only parts of it. In order to improve the overall quality of the system, it is important to detect these errors by assigning a so called confidence measure to each translated word, phrase or sentence. Ideally this measure would be the probability of correctness. An ideal word-level estimator would therefore be the probability that a given word appears at a given position in a sentence; using the notations of Section 1.1 (e_i being the i-th word of sentence $\mathbf{e}$), this is expressed by the following formula:

$$\text{word confidence} = P(correct|i, e_i, \mathbf{f}) \qquad (3)$$

and an ideal sentence-level estimator would be:

$$\text{sentence confidence} = P(correct|\mathbf{e}, \mathbf{f}) \qquad (4)$$

However these probabilities are difficult to estimate accurately; this is why existing approaches rely on approximating them or on computing scores which are supposed to monotonically depend on them.

2.2 State of the Art

Confidence estimation is a common problem in artificial intelligence and information extraction in general (Culotta and McCallum, 2004; Gandrabur et al., 2006). When it comes to natural language processing, it has been intensively studied for automatic speech recognition (Mauclair, 2006; Razik, 2007; Guo et al., 2004). We find in literature (Blatz et al., 2003; Ueffing and Ney, 2004; Ueffing and Ney, 2005; Uhrik and Ward, 1997; Akiba et al., 2004; Duchateau et al., 2002) different ways of approximating the probability of correctness or of calculating scores which are supposed to reflect this probability.

There are three dominating approaches to estimation of word-level confidence measures for machine translation:

- Estimate words posterior probabilities on the n-best list or word-lattice produced by the decoder (the idea is that correct words will appear frequently).
- Use as a confidence estimation the probability that a word in the generated sentence is the translation of a word in the source sentence, by using a translation table.
- Transform each word into a vector of numerical features (for example scores coming from different specialised confidence estimators) and train a perceptron to discriminate between "correct" and "incorrect" classes.

In (Ueffing and Ney, 2004) different original word-level confidence measures are proposed: the word posterior probabilities are estimated from the n-best list, allowing some variation in words positions, and a word's correctness probability is estimated using the translation table generated by an IBM-1 model. Many different confidence measures are investigated in (Blatz et al., 2003). They are based on source and target language models features, n-best lists, words-lattices, or translation tables. The authors also present efficient ways of classifying words or sentences as "correct" or "incorrect" by using naïve Bayes, single- or multi-layer perceptron.

2.3 Our Approach to Confidence Estimation

In the following we will present four different word-level estimators, based on:

- Intra-language mutual information (intra-MI) between words in the generated sentence.
- Inter-language mutual information (inter-MI) between source and target words.
- An n-gram model of the target language.
- A target language model based on linguistic features.

Mutual Information has been proved suitable for building translation tables (Lavecchia et al., 2007)

or alignment models (Moore, 2005). We use intra-language MI to estimate the relevance of a word in the candidate translation given its context (it is supposed to reflect the lexical consistency). Inter-language MI based confidence estimation gives an indication of the relevance of a translation by checking that each word in the hypothesis can indeed be the translation of a word in the source sentence. N-gram and linguistic features models estimate the lexical and grammatical correctness of the hypothesis. Finally, because each of these measures targets a specific kind of error, they can be linearly combined in order to obtain a more powerful confidence measure. The weights are optimised on a development corpus. Each of these estimators produces a score for every word. This score is then compared to a threshold and the word is labelled as "correct" if its score is greater, or "incorrect" otherwise. This classification is then compared to a man made reference which gives an estimation of the efficiency of the measures, in terms of error rate, ROC curve and F-measure (see Section 2.3.1).

2.3.1 Evaluation of the Confidence Measures

As explained before, the CMs are evaluated on a classification task. We manually classified words from 819 sentences generated by MOSES (Koehn et al., 2007) as candidate translations in "French" of English sentences extracted from the test corpus of our system (8067 English words, 8816 French words) and ran our classifiers on the same sentences. A word was classified as correct if its score was above a given threshold. The results were then compared to the human-made references. We used the following metrics to estimate how well our classifier behaved:

Classification Error Rate (CER) is the proportion of errors in classification:

$$\frac{\text{number of incorrectly classified words}}{\text{total number of words}}$$

Correct Acceptance Rate (CAR) is the proportion of correct words retrieved:

$$\frac{\text{number of correctly accepted words}}{\text{total number of correct words}}$$

Correct Rejection Rate (CRR) is the proportion of incorrect words labelled as such:

$$\frac{\text{number of correctly rejected words}}{\text{total number of incorrect words}}$$

F-measure is the harmonic mean of CAR and CRR:

$$F = \frac{2 \times CAR \times CRR}{CAR + CRR}$$

These metrics are common in confidence estimation for machine translation (Blatz et al., 2003). Basically a relaxed classifier has a high CAR (most correct words are labelled as such) and low CRR (many incorrect words are not detected), while a harsh one has a high CRR (an erroneous word is often detected) and a low CAR (many correct words are rejected).
As the acceptance threshold goes from 0 to 1, the classifier becomes harsher: CAR goes from 1 to 0 and CRR from 0 to 1. Therefore we plot CRR against CAR for different thresholds. This tool, very common in information retrieval, is called a *ROC curve*. The ROC curve of a perfect classifier would be the point (1,1) alone, therefore we expect a good classifier to draw a curve as close as possible to the top and right edges of the unit square. This representation is very useful in order to compare the performance of two classifiers: generally speaking, a classifier is better than another if it's ROC curve is always above. In particular, it can be used to quickly visualise the improvement compared to the most naive classifier, which assigns a random score (between 0 and 1) to each word. The ROC curve of such a classifier is the segment going from (0,1) to (1,0), which we plot on our figures. The higher above this line a classifier is, the better. We also plotted on the same diagrams F-measure and CER against CAR.

3 SOFTWARE AND MATERIAL DESCRIPTION

Experiments were run using an English to French phrase-based translation system. We trained a system corresponding to the baseline described in the *ACL workshop on statistical machine translation* (Koehn, 2005). It is based on the state of the art IBM-5 model (Brown et al., 1994) and has been trained on the EUROPARL corpus (proceedings of the European Parliament, (Koehn, 2005)) using GIZA++ (Och, 2000) and the SRILM toolkit (Stolcke, 2002). The decoding process is handled by MOSES. The French vocabulary was composed of 63,508 words and the English one of 48,441 words. We summarise in Table 1 the sizes of the different parts of the corpus. This system achieves state of the art performances.

	sentences pairs	running words	
set		English	French
Learning	465,750	9,411,835	10,211,388
Development	3000	75,964	82,820
Test	1444	14,077	14,705

Table 1: Corpora sizes

4 MUTUAL INFORMATION BASED CONFIDENCE MEASURES

4.1 Mutual Information for Language Modelling

In probability theory mutual information measures how mutually dependent are two random variables. It can be used to detect pairs of words which tends to appear together in sentences. Guo proposes in (Guo et al., 2004) a word-level confidence estimation for speech recognition based on mutual information. In this paper we will compute inter-word mutual information following the approach in (Lavecchia et al., 2007), which has been proved suitable for generating translation tables, rather than Guo's.

$$MI(x,y) = p(x,y)log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (5)$$

$$p(x,y) = \frac{N(x,y)}{N}$$

$$p(x) = \frac{N(x)}{N}$$

where N is the total number of sentences, $N(x)$ is the number of sentences in which x appears and $N(x,y)$ is the number of sentences in which x and y co-occur. We smooth the estimated probability distribution, as in Guo's paper, in order to avoid null probabilities:

$$N(x,y) \leftarrow N(x,y)+C \quad (6)$$

$$p(x,y) \leftarrow \frac{p(x,y)+\alpha p(x)p(y)}{1+\alpha} \quad (7)$$

in which C is a non-negative integer and α a non-negative real number. For example, words like *"ask"* and *"question"* have a high mutual information, while words coming from distinct lexical fields (like *"poetry"* and *"economic"*) would have a very low one. Since it is not possible to store a full matrix in memory, only the most dependent word pairs are kept: we obtain a so called *triggers list*.

4.2 Confidence Measure based on Intra-Language Mutual Information

By estimating which target words are likely to appear together in the same sentence, intra-language MI based confidence score is supposed to reflect the lexical consistency of the generated sentence. We computed mutual information between French words from the French part of the bilingual corpus. Table 2 shows an example of French intra-lingual triggers, sorted by decreasing mutual information.

first word	→	*triggered word*	mutual information
sécurité	→	*alimentaire*	$4.43\cdot10^{-3}$
sécurité	→	*étrangère*	$4.27\cdot10^{-3}$
sécurité	→	*politique*	$4.06\cdot10^{-3}$
	...		
politique	→	*commune*	$1.00\cdot10^{-2}$
politique	→	*économique*	$8.46\cdot10^{-3}$
politique	→	*étrangère*	$7.88\cdot10^{-3}$

Table 2: An example of French intra-lingual triggers

Let $\mathbf{e} = e_1..e_I$ be the target sentence. The score assigned to e_i is the weighted average mutual information between e_i and the words in its context:

$$C(e_i) = \frac{\sum_{j=1..I, j\neq i} w(|i-j|)MI(e_j,e_i)}{\sum_{j=1..I, j\neq i} w(|i-j|)} \quad (8)$$

where $w()$ is a scaling function lowering the importance of long range dependencies. It can be constant if we do not want to take words' positions into account, exponentially decreasing if we want to give more importance to pairs of words close to each other, or a shifted Heaviside function if we want to allow triggering only within a given range (which we will refer to as *triggering window*).
We also experimented with different kinds of normalisation:

- Beforehand normalisation of the **triggers list**:

$$MI(x,y) \leftarrow \frac{MI(x,y)}{max_{y'}MI(x,y')} \quad (9)$$

- Normalisation with regard to **norm**-1 as in (Ueffing and Ney, 2004):

$$C(e_i) \leftarrow \frac{C(e_i)}{\sum_{j=1}^{I}|C(e_j)|} \quad (10)$$

- With regard to **norm**-∞:

$$C(e_i) \leftarrow \frac{C(e_i)}{\max_{j=1..I}|C(e_j)|} \quad (11)$$

Tool words like *"the"*, *"of"*,... tend to have a very high mutual information with all other words thus polluting the trigger list. We therefore ignored them in some of our experiments.

Presenting the performances of the confidence measure with all different settings (normalisation, with or without tool words,...) would be tedious. Therefore we only show the settings that yield the best performances. Note that while other settings often yield much worse performance, a few perform

almost as well, therefore there are no definite "optimal settings". Figure 1 shows the ROC curve, CER and F-measure of a classifier based on intra-MI: tool words were ignored, no normalisation was applied, and words positions were not taken into account. Remember that these curves are obtained by assigning a score to each word in the generated sentences, then, for different thresholds between 0 and 1, classifying all these words as correct or incorrect. Each of these thresholds gives a CAR, a CRR and a CER and therefore a point of the curves.

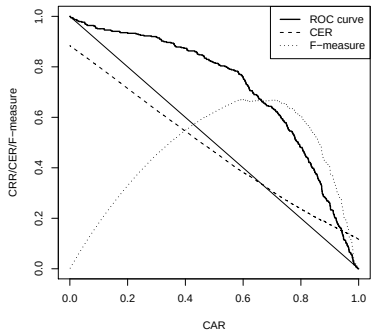

Figure 1: Intra-MI, no tool words, no normalisation, no weighting nor triggering window.

This classifier shows very interesting discriminating power : for a CAR of 50% the CRR is slightly above 80% (harsh classifier), and for a CRR of 50% the CAR is almost 80% (relaxed classifier). We empirically found that taking word positions into account in intra-MI based confidence measures tends to yield lower performance. We interpret in the following way: intra-language MI reflects lexical consistency of the sentence, but two related words may not be next to each other in the sentence.

4.3 Confidence Measure Based on Inter-language Mutual Information

The principle of intra-language MI was to detect which words trigger the appearance of an other word in the same sentence. This principle can be extended to pairs of source and target sentences (Lavecchia et al., 2007): let $N_S(x)$ be the number of source sentences in which x appears, $N_T(y)$ the number of target sentences in which y appears, $N(x,y)$ the number of pairs *(source sentence, target sentence)* such that x appears in the source and y in the target, and N the total number of pairs of source and target sentences. Then let us define:

$$\begin{aligned} p_S(x) &= \frac{N_S(x)}{N} \\ p_T(y) &= \frac{N_T(y)}{N} \\ p(x,y) &= \frac{N(x,y)}{N} \\ MI(x,y) &= p(x,y) log_2 \left(\frac{p(x,y)}{p_S(x) p_T(y)} \right) \end{aligned} \quad (12)$$

Guo's smoothing can be applied as in Section 4.2. One then keeps only the best triggers and obtain a so-called *inter-lingual triggers list*. Table 3 shows an example of such triggers between English and French words, sorted by decreasing mutual information.

English word	→	*triggered French word*	MI
security	→	*sécurité*	$8.03 \cdot 10^{-2}$
security	→	*étrangère*	$8.55 \cdot 10^{-3}$
security	→	*politique*	$6.08 \cdot 10^{-3}$
	...		
policy	→	*politique*	$2.62 \cdot 10^{-2}$
policy	→	*commune*	$3.39 \cdot 10^{-3}$
policy	→	*étrangère*	$2.71 \cdot 10^{-3}$

Table 3: An Example of Inter-Lingual triggers

The confidence measure is then:

$$C(e_i) = \frac{\sum_{j=1}^{J} w(|i-j|) MI(f_j, e_i)}{\sum_{j=1}^{J} w(|i-j|)} \quad (13)$$

We show in Figure 2 the characteristics of such an inter-MI based classifiers. No normalisation whatsoever was applied, and tool words were excluded. This time triggering was allowed within a window of width 9 centred on the word the confidence of which was being evaluated.

Unlike intra-MI based classifier, we found here that setting a triggering window yields the best performance. This is because inter-language MI indicates which target words are possible translations of a source word. This is much stronger than the lexical relationship indicated by intra-MI; therefore allowing triggering only within a given window or simply giving less weight to "distant" words pairs reflects the fact that words in the source sentence and their translations in the target sentence appear more or less in the same order (this is the same as limiting the distorsion, which is the difference between the positions of a word and its translation).

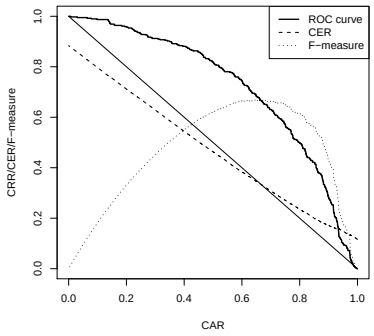

Figure 2: Inter-language MI based CM: tool words excluded, no normalisation, triggering is allowed within a centred window of width 9.

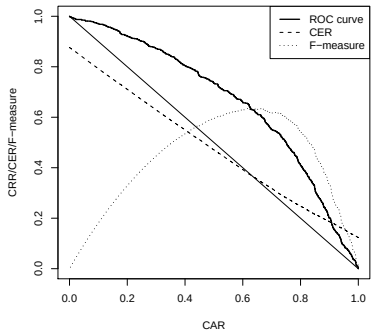

Figure 3: Performance of a 4-grams language model based classifier.

5 N-GRAMS BASED CONFIDENCE MEASURE

5.1 Principle

Remember Equation 2: the decoder makes a compromise between $P(\mathbf{e})$ (which we will refer to as *language model score*) and $P(\mathbf{f}|\mathbf{e})$ (*translation score*). Because of that, if a candidate **e** has a high translation score and a low language model score, it might be accepted as the "best" translation. But a low LM score often means an incorrect sentence and therefore a bad translation. This consideration applies on subsentence level as well as on sentence level: if the n-gram probability of a word is low, it often means that it is wrong or at least misplaced. Therefore we want to use the language model alone in order to detect incorrect words. We decided to use the word probability derived from an n-gram model as a confidence measure:

$$C(e_i) \quad = \quad P(e_i|e_{i-1},..,e_{i-n+1}) \qquad (14)$$

While intra-language triggers are designed to estimate the lexical consistency of the sentence, this measure is supposed to estimate its well-formedness. Figure 3 shows the performances of a 4-grams based classifier.

While still showing an interesting discriminating power, it does not perform as well as the best MI-based classifiers: some hypothesis with a low language model score will indeed have already been discarded by the decoder. Also we classify as incorrect only the last word of the n-gram, however a low n-gram score indicates that the sequence (or any word in it) is wrong, rather than only the last word.

6 LINGUISTIC FEATURES BASED CONFIDENCE MEASURE

6.1 Principle

Classical language models do not directly take into account tense, gender and number agreement between the different words of the output sentence. We want to specifically target agreement errors: this is why in the following we propose a confidence measure based on linguistic features. For that, we use BDLEX (De Calmès and Pérennou, 1998) to replace each word by a vector of features (Smaïli et al., 2004). We specifically select three features:

- **Syntactic class**, for example noun, verb, etc...
- **Tense** of verbs or nothing for other classes.
- **Number and gender** of nouns, adjectives or past participles, **person** for verbs, nothing otherwise.

For example the word *était* becomes *V,ii,3S* standing for "verb, imperfect indicative, 3rd person". We then train a classical n-gram model on the generated features corpus using the SRILM toolkit and use the n-gram probability as a confidence estimation. In Figure 4 we print the ROC, CER and F-measures of confidence measures based on a 6-grams linguistic features language model.

The performances of this CM are rather disappointing, and the CER is particularly terrible. This can probably be at least partially explained by the difficulty of disambiguation (some words belong to different classes, like the French word "sommes" which can be a conjugated verb or a plural noun): because we have no information that might allow us to perform a correct choice, it was randomly performed during training of the model, and during sentence evaluation the most likely class (according to the previously

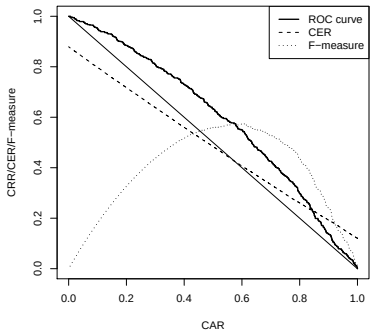

Figure 4: Performance of a classifier based on a 6-grams linguistic features model.

trained linguistic feature n-gram model) was chosen. We believe progress could certainly be made by performing smarter disambiguation.

7 FUSION OF CONFIDENCE MEASURES

We linearly combined the scores assigned to each word by different confidence measures to produce a new score. The weights are optimised by the perceptron algorithm on a small corpus (600 pairs of sentences), and tested on a different corpus (219 pairs of sentences). Figure 5 shows the performances of the classifier resulting from the linear combination of the best previously presented intra-MI based classifier and the best inter-MI based one.

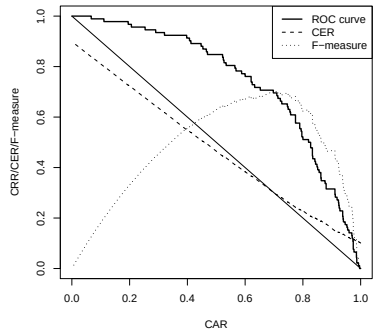

Figure 5: combination of the two best MI based CMs.

The combination of the two yields interesting improvement, especially in terms of error rate. The perceptron gave more weights to the inter-MI based scores, but that is because these scores are generally lower and does not mean that this measure is more significant. On the other hand, combining these two confidence measures with n-grams and linguistic features based ones did not bring any improvement over our test corpus.

8 DISCUSSION AND CONCLUSION

In this article, we present confidence scores that showed interesting discriminating power. We summarised the results obtained by the best different estimators (in terms of F-measure) in Table 4. For comparison Blatz et al. obtain in (Blatz et al., 2003) a CER of 29.2% by combining two different word posterior probability estimates (with and without alignment) and the translation probabilities from IBM-1 model.

	threshold	CER	CAR	CRR	F-measure
intra-MI	$3.6 \cdot 10^{-5}$	0.383	0.600	0.760	0.700
inter-MI	0.0008	0.368	0.620	0.724	0.668
4-gram model	0.134	0.377	0.619	0.653	0.636
linguistic 6-grams	0.188	0.422	0.578	0.574	0.576
combined MI	$9 \cdot 10^{-5}$	0.251	0.759	0.663	0.708

Table 4: Performances of the best classifiers.

The settings used by the best intra-MI based confidence measures were the following: tool words were ignored, no normalisation was applied, and words positions were not taken into account. For the best inter-MI based CMs, tool words were not taken into account, no normalisation, and triggering was allowed within a centred window of width 9 (maximal "distortion" of 4). From these figures we can tell that the best MI-based confidence measures outperform significantly the other CMs presented here, especially when used in combination. Note however that the best classifiers in terms of F-measure are not necessarily the best ones with regard to other metrics, for example CER.

8.1 Application of Confidence Measures

While they were not investigated in this article, we can imagine several applications to confidence estimation, beside manual correction of erroneous words: **pruning or reranking of the n-best list** according to the confidence score, **generation of new hypothesis** by recombining parts of different candidates having high scores, or **discriminating training** by tuning the parameters to optimise the separation between sentences (or words, or phrases) having a high confidence score (hopefully they are correct translations) and sentences having a low one.

8.2 Prospects

We plan to go further in our investigation on confidence measures for SMT: first, while the confidence measures we used take into account word insertion and word substitution, they do not directly take into account word deletion nor word order, and neither do our reference corpus (in which words are labelled as correct or incorrect, but missing words are not indicated). This serious drawback has to be addressed. Assigning confidence scores to alignment might help to this end. Second, we believe that in a context of phrase-based translation, phrase-level confidence estimation would be more appropriate. Also many features used in speech recognition or automatic translation could be used in confidence estimation: distant models, word alignment, word spotting, etc... Another problem is the fusion of different classifiers. We use a very simple single layer perceptron, but many solutions have been proposed in literature to achieve more appropriate merging. Finally, progress could be made on classifiers' evaluation: because classifying a word as correct or incorrect is a very difficult task even for a human translator, and because the results of such a task may vary according to the translator or worse, vary along time for a given translator, we should combine different human-generated references.

REFERENCES

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

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2003). Confidence estimation for machine translation. final report, jhu/clsp summer workshop.

Brown, P., Pietra, S., Pietra, V., and Mercer, R. (1994). The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Culotta, A. and McCallum, A. (2004). Confidence estimation for information extraction. *Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

De Calmès, M. and Pérennou, G. (1998). Bdlex: a lexicon for spoken and written french. In *Proceedings of 1st International Conference on Langage Resources & Evaluation*.

Duchateau, J., Demuynck, K., and Wambacq, P. (2002). Confidence scoring based on backward language models. *Acoustics, Speech, and Signal Processing, 2002. Proceedings.(ICASSP'02). IEEE International Conference on*, 1.

Gandrabur, S., Foster, G., and Lapalme, G. (2006). Confidence estimation for NLP applications. *ACM Transactions on Speech and Language Processing*, 3(3):1–29.

Guo, G., Huang, C., Jiang, H., and Wang, R. (2004). A comparative study on various confidence measures in large vocabulary speech recognition. *2004 International Symposium on Chinese Spoken Language Processing*, pages 9–12.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT Summit*, 5.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstation session*.

Lavecchia, C., Smaili, K., Langlois, D., and Haton, J. (2007). Using inter-lingual triggers for machine translation. *Eighth conference INTERSPEECH*.

Mauclair, J. (2006). *Mesures de confiance en traitement automatique de la parole et applications*. PhD thesis, LIUM, Le Mans, France.

Moore, R. C. (2005). Association-based bilingual word alignment. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts, Ann Arbor, Michigan, pp. 1-8*.

Och, F. (2000). Giza++ tools for training statistical translation models.

Razik, J. (2007). *Mesures de Confiance trame-synchrones et locales en reconnaissance automatique de la parole*. PhD thesis, LORIA, Nancy, FRANCE.

Smaïli, K., Jamoussi, S., Langlois, D., and Haton, J. (2004). Statistical feature language model. *Proc. ICSLP*.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. pages 901–904.

Ueffing, N. and Ney, H. (2004). Bayes decision rule and confidence measures for statistical machine translation. pages 70–81. Springer.

Ueffing, N. and Ney, H. (2005). Word-level confidence estimation for machine translation using phrase-based translation models. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 763–770.

Uhrik, C. and Ward, W. (1997). Confidence Metrics Based on N-Gram Language Model Backoff Behaviors. In *Fifth European Conference on Speech Communication and Technology*. ISCA.