Word confidence estimation for speech translation

Laurent Besacier, Benjamin Lecouteux, Ngoc-Quang Luong, K Hour, Marwa Hadj Salah

WORD CONFIDENCE ESTIMATION FOR SPEECH TRANSLATION

L. Besacier, B. Lecouteux, N.Q. Luong, K. Hour and M. Hadjsalah

LIG, University of Grenoble, France

laurent.besacier@imag.fr

Abstract

Word Confidence Estimation (WCE) for machine translation (MT) or automatic speech recognition (ASR) consists in judging each word in the (MT or ASR) hypothesis as correct or incorrect by tagging it with an appropriate label. In the past, this task has been treated separately in ASR or MT contexts and we propose here a joint estimation of word confidence for a spoken language translation (SLT) task involving both ASR and MT. This research work is possible because we built a specific corpus which is first presented. This corpus contains 2643 speech utterances for which a quintuplet containing: ASR output (src-asr), verbatim transcript (src-ref), text translation output (tgt-mt), speech translation output (tgt-slt) and post-edition of translation (tgt-pe), is made available. The rest of the paper illustrates how such a corpus (made available to the research community) can be used for evaluating word confidence estimators in ASR, MT or SLT scenarios. WCE for SLT could help rescoring SLT output graphs, improving translators productivity (for translation of lectures or movie subtitling) or it could be useful in interactive speech-to-speech translation scenarios.

Word confidence estimation (WCE), Spoken Language Translation (SLT), Corpus, Joint features.

1. Introduction

Confidence estimation is a rather hot topic both for Automatic Speech Recognition (ASR) and for Machine Translation (MT). While ASR and MT systems produce more and more user-acceptable outputs, we still face open questions such as: are these translations/transcripts ready to be published as they are? Are they worth to be corrected or do they require retranslation/retranscription from scratch? It is undoubtedly that building a method which is capable of pointing out the correct parts as well as detecting the errors in each MT or ASR hypothesis is crucial to tackle these above issues. Also, confidence estimation can help to re-rank N-best hypotheses [1] or re-decode the search graph [2]. If we limit the concept "parts" to "words", the problem is called Word-level Confidence Estimation (WCE).

The WCE's objective is to assign each word in the MT or ASR hypothesis a confidence score (typically between 0 and 1). For error detection, this score can be binarized and then each word is tagged as correct or incorrect. In that case, a classifier which has been trained beforehand from a feature set calculates the confidence score for the output word, and then compares it with a pre-defined threshold. All words with scores that exceed this threshold are categorized in the *Good* label set; the rest belongs to the *Bad* label set. In the past, this task has been treated separately in ASR or MT contexts and we propose here a joint estimation of word confidence for a spoken language translation (SLT) task involving both ASR and MT. We believe that WCE for SLT could help improving translators productivity (for lecture or movie translation) or it could be useful in interactive speech-to-speech translation.

The remaining of this paper is the following. In section 2 we present our first contribution: a corpus (distributed to the research community) dedicated to WCE for SLT. To our knowledge, this is the first corpus that allows experimenting such a task. It contains 2643 speech utterances for which a quintuplet (containing ASR output, verbatim transcript, text translation output, speech translation output and post-edition of translation) is available. Then sections 3 and 4 present our WCE systems (as well as a quick description of related works) for ASR and MT respectively. Section 5 illustrates how our corpus can be used for evaluating word confidence estimators in a SLT scenario. Finally we conclude this paper and give some perspectives.

2. A database for WCE evaluation in spoken language translation

2.1. Starting point: an existing MT Post-edition corpus

For a French-English translation task, we used our SMT system to obtain the translation hypothesis for 10,881 source sentences taken from news corpora of the WMT (Workshop on Machine Translation) evaluation campaign (from 2006 to 2010). Post-editions were obtained from non professional translators using a crowdsourcing platform. More details on the baseline SMT system used can be found in [4] and more details on the post-edited corpus can be found in [5]. It is worth mentionning, however, that a subset (311 sentences) of these collected post-editions was assessed by a professional translator and 87.1% of post-editions were judged to improve the hypothesis

Then, the word label setting for WCE was done using TERp-A toolkit [3]. Table 1 illustrates the labels generated by TERp-A for one hypothesis and post-edition pair. Each word or phrase in the hypothesis is aligned to a word or phrase in the post-edition with different types of edit: I (in-

Reference	The	consequence	of	the	fundamentalist	movement		also	has	its importance	.
		S			S	Y	I		D	P	
Hyp After Shift	The	result	of	the	hard-line	trend	is	also		important	.

Table 1: Example of WCE label setting using TERp-A [3]

sertions), S (substitutions), T (stem matches), Y (synonym matches), and P (phrasal substitutions). The lack of a symbol indicates an exact match and will be replaced by E thereafter. We do not consider the words marked with D (deletions) since they appear only in the reference. However, later on, we will have to train binary classifiers (good/bad) so we re-categorize the obtained 6-label set into binary set: The E, T and Y belong to the *Good* (G), whereas the S, P and I belong to the *Bad* (B) category. Finally, we observed in our corpus that out of total words (train and test sets) are 85% labeled G, 15% labeled B.

From this corpus, we extract 10,000 triplets (source reference *src-ref*, machine translation output *tgt-mt* and post-edition of translation *tgt-pe*) for training our WCE (for MT) system and keep the remaining 881 triplets as a test set.

2.2. Augmenting the corpus with speech recordings and transcripts

In order to take advantage of the existing PE corpus, we decided to record the utterances of its test part to augment the corpus with speech inputs. We admit that this would have been better to capture real speech data, then transcribe it, translate and post-edit but we believe that our corpus will remain useful to study WCE for SLT, even if translating read speech is not the best practical SLT task we could imagine.

So, the test set of this corpus was recorded by French native speakers. Each of the 881 sentences was uttered by 3 speakers, leading to 2643 speech recordings. 15 speakers (9 women and 6 men) took part to the speech data collection in normal office condition. The total length of the speech corpus obtained is more than 5h since some utterances were pretty long.

Then, our French ASR system based on KALDI toolkit [6] was used to obtain the speech transcripts. The 3-gram language model was trained on the French ESTER corpus as well as French Gigaword (vocabulary size is 55k). SGMM-based acoustic models were trained using the same ESTER corpus - see details in [7].

It is important to note that automatic post-processing was needed at the output of the ASR system in order to match requirements of standard input for machine translation (we wanted our ASR outputs to match, as much as possible, our already available *src-ref* utterances). Thus, the following post-treatments were applied: number conversion (back to digit numbers), recasing (our SMT system is a true case one), re-punctuating, converting full words back to abbreviations (*kilometre* becomes *km*, *madame* becomes *Mme*, etc.) and restoring special characters (*pourcents* becomes %, *euro* becomes €). With this post-processing, the output of our ASR system, scored against the *src-ref* reference went from 29.05% WER to 26.6% WER.

This WER may appear as rather high according to the task (transcribing read news) but these news contain a lot of foreign named entities (part of the data is extracted from French newspapers dealing with european economy in many EU countries).

2.3. Obtaining labels in order to evaluate WCE for SLT

We now have a new element of our desired quintuplet: the ASR output *src-asr*. It is the noisy version of our already available verbatim transcripts called *src-ref*. This ASR output (*src-asr*) was then translated by the exact same SMT system [4] already mentionned in paragraph 2.1. This new output translation is called *tgt-slt* and it is a degraded version of *tgt-mt*.

At this point, a strong assumption we made has to be revealed: we re-used the post-editions obtained from the text translation task (called *tgt-pe*), to infer the quality (G,B) labels of our speech translation output *tgt-slt*. The word label setting for WCE is also done using TERp-A toolkit [3] between *tgt-slt* and *tgt-pe*. This assumption (as well as the fact that initial MT post-edition can be also used to infer labels of a SLT task) is reasonable regarding results (later presented in Table 4) where it is shown that there is not a huge difference between the MT and SLT performance (evaluated with BLEU). This means that if the real SLT output had been post-edited, we would have obtained very similar PE to the actual ones.

The remark above is important and this is what makes the value of this corpus. For instance, other corpora such as the TED corpus compiled by LIUM[1] contains also a quintuplet with ASR output, verbatim transcript, MT output, SLT output and target translation. But there are two main differences: first, the target translation is a manual translation of the prior subtitles so this is not a post-edition of an automatic translation (and we have no guarantee that the G/B labels extracted from this will be reliable for WCE training and testing). Secondly, in our corpus, each sentence is uttered by 3 different speakers in order to introduce a minimum of speaker variability in the test set (the consequence is that we have different ASR outputs for a single source sentence).

2.4. Final corpus statistics and web link for download

The main statistics regarding this corpus are in Table 2, where we also clarify how the WCE labels were obtained. For the test set, we now have all the data needed to evaluate WCE for 3 tasks :

[1] http://www-lium.univ-lemans.fr/fr/content/corpus-ted-lium

- **ASR**: extract G/B labels by computing WER between *src-asr* and *src-ref*,
- **MT**: extract G/B labels by computing TERp-A between *tgt-mt* and *tgt-pe*,
- **SLT**: extract G/B labels by computing TERp-A between *tgt-slt* and *tgt-pe*.

Data	# train utt	# test utt	method to obtain WCE labels
src-ref	10000	881	
src-sig		5h	speech
src-asr		881*3	wer(*src-asr*,*src-ref*)
tgt-mt	10000	881	terpa(*tgt-mt*,*tgt-pe*)
tgt-slt		881*3	terpa(*tgt-slt*,*tgt-pe*)
tgt-pe	10000	881	

Table 2: Overview of our post-edition corpus for SLT

Table 3 gives an example of quintuplet available in our corpus. One transcript (*src-hyp1*) has 1 error while the other one (*src-hyp2*) has 4. This leads to respectively 2 B labels (*tgt-slt1*) and 4 B labels (*tgt-slt2*) in the speech translation output, while *tgt-mt* has only one B label. Table 4 summarizes the MT (translation from verbatim transcripts) and SLT (translation from automatic speech transcripts) performances obtained on our corpus, as well as the distribution of good (G) and bad (B) labels inferred for both tasks. Logically, the percentage of (B) labels increases from MT to SLT task in the same conditions.

src-ref	quand	notre	cerveau	chauffe
src-hyp1	*comme*	notre	cerveau	chauffe
labels ASR	B	G	G	G
src-hyp2	*qu'*	*entre*	*serbes*	*au* chauffe
labels ASR	B	B	B	B G
tgt-mt	when	our	brains	*chauffe*
labels MT	G	G	G	B
tgt-slt1	*as*	our	brains	*chauffe*
labels SLT	B	G	G	B
tgt-slt2	*between*	*serbs*	*in*	*chauffe*
labels SLT	B	B	B	B
tgt-pe	when	our	brain	heats up

Table 3: Exemple of quintuplet with associated labels

task	ASR (WER)	MT (BLEU)	% G (good)	% B (bad)
tgt-mt	0%	36.1%	82.5%	17.5%
tgt-slt	26.6%	30.6%	65.5%	34.5%

Table 4: MT and SLT performances on our test set

This corpus is available for download on *github.com/besacier/WCE-SLT-LIG*.

3. WCE for speech transcription

3.1. Related work

Several previous works tried to propose effective confidence measures in order to detect errors on ASR outputs. Out-Of-Vocabulary (OOV) detection was introduced by [8] and extended by [9] and [10]. [9] introduced the use of word posterior probability (WPP) as a confidence measure for speech recognition. Posterior probability of a word (or a sequence) is most of the time computed using the hypothesis word graph [9] [11].

Recent approaches [12, 10] for confidence measure estimation use side-information extracted from the recognizer: normalized likelihoods (WPP), the number of competitors at the end of a word (hypothesis density), decoding process behavior, linguistics features, acoustic features (acoustic stability, duration features) and semantic features. Finally, these papers show the prominence of linguistic features.

Later, WPP score was combined with other high-level knowledge sources to improve the confidence estimation. For instance, [10] proposed an efficient method that combines various features (acoustic, linguistic, decoding and semantic features). Another work by [13] combines scores extracted from several sources: N-best features, acoustic stability, hypothesis density, duration features, language model, parsing features, WPP, etc.

3.2. WCE system used and baseline performance

In this work, we extract several types of features, which come from the ASR graph, from language model scores and from a morphosyntactic analysis. These features are listed below:

- Acoustic features : words errors probably induce acoustic distortions between the hypothesis and the best phonetic sequence. Many observations points out that word length can predict correct words and errors: we add a feature which consists of the word duration (F-dur).
- Graph features : they are extracted from the word confusion networks. When an error occurs, the search algorithm explores various alternative paths: the posterior probabilities and alternative paths can help to predict errors. We use the number of alternative (F-alt) paths in the word section, and the posterior probability (F-post).
- Linguistic features : they are based on probabilities provided by the language model (3-gram LM) used in the KALDI ASR system. We use the word itself (F-word) and the 3-gram probability (F-3g) . We also add the feature (F-back), proposed in [12] which represents the back-off level of the targeted word.
- Lexical Features: word's Part-Of-Speech (F-POS) are computed using tree-tagger for French.

We use a variant of boosting classification algorithm in order to combine features. The used implementation is Bonzaiboost[2] [14]. It implements the boosting algorithm Adaboost.MH over deeper trees.

For each word, we estimate the 7 features (F-Word; F-3g; F-back; F-alt; F-post; F-dur; F-post) previously described. The classifier is trained on BREF 120 corpus [15]. After

[2]http://bonzaiboost.gforge.inria.fr

decoding, we obtain about 1M word examples. Each word from this corpus is tagged as correct or not correct, according to the reference.

Once we have the prediction model built with all features, we apply it on the test set (3*881 sentences) and obtained the required WCE labels along with confidence probabilities. In term of F-score, our WCE system reaches the following performance: predicting *"G"* label: (**87.85%**), and predicting *"B"* label: (**37.28%**).

4. WCE for machine translation

4.1. Related work

The Workshop on Machine Translation (WMT) introduced in 2013 a WCE task for machine translation. [16, 17] employed the Conditional Random Fields (CRF) [18] model as their Machine Learning method to address the problem as a sequence labeling task. Meanwhile, [19] extended the global learning model by dynamic training with adaptive weight updates in the perceptron training algorithm. As far as prediction indicators are concerned, [19] proposed seven word feature types and found among them the "common cover links" (the links that point from the leaf node containing this word to other leaf nodes in the same subtree of the syntactic tree) the most outstanding. [16] focused only on various n-gram combinations of target words. Inheriting most of previously-recognized features, [17] integrated a number of new indicators relying on graph topology, pseudo reference, syntactic behavior (constituent label, distance to the semantic tree root) and polysemy characteristic. Optimization endeavors were also made to enhance the baseline, including classification threshold tuning, feature selection and boosting technique [17].

4.2. WCE system used and baseline performance

We employ the Conditional Random Fields [18] (CRFs) as our machine learning method, with WAPITI toolkit [20], to train the WCE model. A number of knowledge sources are employed for extracting features, in a total of 25 major feature types:

- Target Side: target word; bigram (trigram) backward sequences; number of occurrences
- Source Side: source word(s) aligned to the target word
- Alignment Context [21]: the combinations of the target (source) word and all aligned source (target) words in the window ± 2
- Word posterior probability [22]
- Pseudo-reference (Google Translate): Does the word appear in the pseudo reference or not?
- Graph topology [23]: number of alternative paths in the confusion set, maximum and minimum values of posterior probability distribution
- Language model (LM) based: length of the longest sequence of the current word and its previous ones in the target (resp. source) LM. For example, with the target word w_i: if the sequence $w_{i-2}w_{i-1}w_i$ appears in the target LM but the sequence $w_{i-3}w_{i-2}w_{i-1}w_i$ does not, the n-gram value for w_i will be 3.
- Lexical Features: word's Part-Of-Speech (POS); sequence of POS of all its aligned source words; POS bigram (trigram) backward sequences; punctuation; proper name; numerical
- Syntactic Features: null link [24]; constituent label; depth in the constituent tree
- Semantic Features: number of word senses in WordNet.

Interestingly, this feature set was also used in our English - Spanish WCE System submitted for WMT 2013 Quality Estimation shared task and obtained the best performance [23].

Once we have the prediction model, we apply it on the test set (881 sentences) and obtained the required WCE labels along with confidence probabilities. In term of F-score, our WCE system reaches very promising performance in predicting *"G"* label (**87.65%**), and acceptable for *"B"* label (**42.29%**).

5. Joint estimation of word confidence for a speech translation task

Now, if we consider WCE for a speech translation task, there is no related work available since, to our knowledge, this is the first time such a task is proposed with a corpus allowing to evaluate joint WCE features coming from both ASR and MT.

task feat. type	**WCE for ASR** ASR feat.	**WCE for MT** MT feat.	**WCE for SLT** MT feat.	**WCE for SLT** ASR feat.	**WCE for SLT** 0.5MT+0.5ASR feat.
F(G)	87.85%	87.65%	77.17%	76.41%	77.54%
F(B)	37.28%	42.29%	39.34%	38.00%	43.96%

Table 5: Summary of word confidence estimation (WCE) results obtained on our corpus with different feature sets based on ASR, MT or both. Numbers reported are F scores for Good (G) and Bad (B) labels respectively with a common decision threshold.

We first report in Table 5 the baseline results by individual WCE systems for a single ASR task and for a single MT task (second and third columns of the table - numbers correspond to the performance of the systems described in the two previous sections). Then, to illustrate how our corpus can be used for word confidence estimation in speech translation, we evaluated the performance of 3 systems (using *labels SLT* - see Table 3 - as reference to score the WCE systems):

- The first system (SLT sys. / MT feat.) is the one described in section 4 and uses only MT features. No

modification of the WCE (for MT) system is needed since the only difference is that the source sentence is *src-hyp* (ASR output) instead of *src-ref*,

- The second system (SLT sys. / ASR feat.) is the one described in section 3 and uses only ASR features. So this is predicting SLT output confidence using only ASR confidence features ! Word alignment information between *src-hyp* and *tgt-slt* is needed to project the WCE scores coming from ASR, to the SLT output (done using adequate Moses option, where the alignment information is kept in the decoding output).
- The third system (SLT sys. / MT+ASR feat.) combines the information from the two previous WCE systems. In this work, the ASR-based confidence score of the source is projected to the target SLT output and linearly combined with the MT-based confidence score (we tried different weights but only report *0.5MT+0.5ASR* as well as *0.9MT+0.1ASR* in the results). It is important to note that WCE systems are not retrained here since we perform a late fusion of scores from two different systems. Training a specific WCE system for SLT based on joint ASR and MT features is part of future work.

The results of these 3 systems are given in the last 3 columns of Table 5. They are obtained on the whole test set [3]. For the late fusion (MT+ASR), we do an arithmetic mean of both WCE systems scores [4]. From these results, we see that the use of both ASR-based and MT-based confidence scores improve the F-score for *"B"* label from 39.34% (MT only features) and 38% (ASR only features) to 43.96% (MT+ASR features), while giving similar F-score for *"G"* label. It is also interesting to notice that using ASR features lead to reasonable performance, almost equivalent to the MT features baseline. This can appear as rather disturbing because in that case, WCE estimator do not look at the translation to predict the confidence of the target words ; it only uses (detected) ASR errors to decide which word is good or bad in the speech translation output.

Figure 1 reports more detailed experiments where the G/B decision threshold varies systematically from 0.5 to 0.9 (with a step of 0.025). The different systems use different linear combination weights.

- *Weight=1* corresponds to the use of MT features only,
- *Weight=0.9* linearly combines both confidence scores as follows: 0.9MT+0.1ASR (intuitively, we thought that MT features would be more important),
- *Weight=0.5* linearly combines both confidence scores as follows: 0.5MT+0.5ASR,
- *Weight=0* corresponds to the use of ASR features only,

[3]They are given to illustrate how our database can be used, with basic strategies to fuse ASR and MT scores. More advanced fusion together with a crossvalidation protocol will be presented in future work

[4]All the results of the table are given using a G/B decision thershold which is a priori set to 0.7

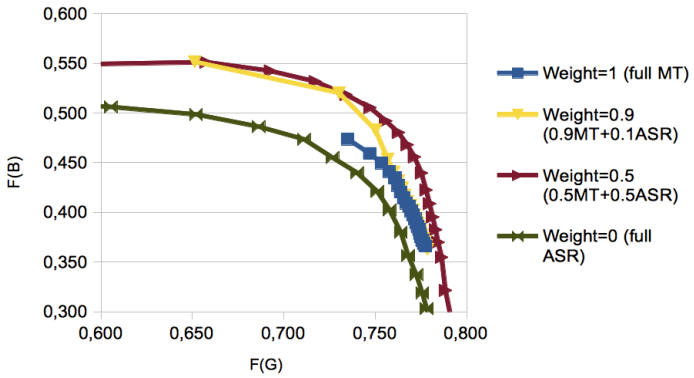

Figure 1: WCE performance (F(B) vs F(G) of different WCE methods - for SLT - for different decision thresholds varying from 0.5 to 0.9).

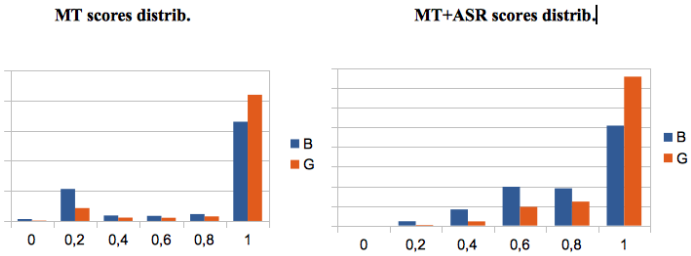

Figure 2: Evolution of the WCE scores distribution from MT features to MT+ASR features

From this figure, we see clearly that using both MT and ASR confidence scores improves the overall WCE performance. However, looking at the results obtained separately by the individual systems, one would have expected a better improvement with their combination. One explanation for this is the fact our WCE scores distributions are rather biased (as seen in Figure 2, many scores equal 1 for both G and B labels). Even if averaging (or linearly combining) ASR and MT scores tend to improve the class separability (Figure 2 shows how the WCE scores distributions evolve from *MT* to *MT+ASR* features), a better strategy might be to replace linear combination by more advanced strategies such as decision trees, SVMs or joint classifier based on the union of ASR and MT features, etc.

6. Conclusion

We presented a specific corpus to study and evaluate word confidence estimation of speech translation. It contains 2643 speech utterances with a quintuplet containing ASR output, verbatim transcript, MT output, SLT output and post-edition of translations. Researchers interested in making use of the dataset can download it from *github.com/besacier/WCE-SLT-LIG*. We also intend to record speech for the 10000 sentences of the train part described in Table 2. The perspectives of this work are numerous:

- propose a new shared task on word confiedence estimation for speech translation,
- train a single WCE system for SLT using joint ASR+MT features and see if more SLT errors can be accurately detected,
- rescore speech translation N-best lists or redecode speech translation graphs using WCE information, as was done by [2] but for MT only,
- use WCE for data augmentation from un-transcribed (and/or un-translated) speech in semi-supervised SLT scenarios,
- adapt WCE system for real interactive speech translation scenarios such as news or lectures subtitling,
- move from a binary (Good or Bad translation) to a 3-class decision problem (Good, ASR error, MT error),
- study how WCE can be adapted to a simultaneous interpetation task.

7. References

[1] N.-Q. Luong, L. Besacier, and B. Lecouteux, "Word Confidence Estimation for SMT N-best List Reranking," in *Proceedings of the Workshop on Humans and Computer-assisted Translation (HaCaT) during EACL*, Gothenburg, Suède, 2014. [Online]. Available: http://hal.inria.fr/hal-00953719

[2] N. Q. Luong, L. Besacier, and B. Lecouteux, "An Efficient Two-Pass Decoder for SMT Using Word Confidence Estimation," in *European Association for Machine Translation (EAMT)*, Dubrovnik, Croatie, jun 2014. [Online]. Available: http://hal.inria.fr/hal-01002922

[3] M. Snover, N. Madnani, B. Dorr, and R. Schwartz, "Terp system description," in *MetricsMATR workshop at AMTA*, 2008.

[4] M. Potet, L. Besacier, and H. Blanchon, "The lig machine translation system for wmt 2010," in *Proceedings of the joint fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT2010)*, A. Workshop, Ed., Uppsala, Sweden, 11-17 July 2010.

[5] M. Potet, R. Emmanuelle E, L. Besacier, and H. Blanchon, "Collection of a large database of french-english smt output corrections," in *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, May 2012.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[7] S. Galliano, E. Geoffrois, G. Gravier, J. F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news," in *In Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006*, 2006, pp. 315–320.

[8] A. Asadi, R. Schwartz, and J. Makhoul, "Automatic detection of new words in a large vocabulary continuous speech recognition system," *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 1990.

[9] S. R. Young, "Recognition confidence measures: Detection of misrecognitions and out-of-vocabulary words," *Proc. of International Conference on Acoustics, Speech and Signal Processing*, pp. 21–24, 1994.

[10] B. Lecouteux, G. Linarès, and B. Favre, "Combined low level and high level features for out-of-vocabulary word detection," *INTERSPEECH*, 2009.

[11] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," *Proc. of European Conference on Speech Communication Technology*, pp. 827–830, 1997.

[12] J. Fayolle, F. Moreau, C. Raymond, G. Gravier, and P. Gros, "Crf-based combination of contextual features to improve a posteriori word-level confidence measures." in *Interspeech*, 2010.

[13] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, 2004.

[14] C. R. Antoine Laurent, Nathalie Camelin, "Boosting bonsai trees for efficient features combination : application to speaker role identification," in *Interspeech*, 2014.

[15] L. F. Lamel, J.-L. Gauvain, M. Eskénazi, *et al.*, "Bref, a large vocabulary spoken corpus for french1," *training*, vol. 22, no. 28, p. 50, 1991.

[16] A. L.-F. Han, Y. Lu, D. F. Wong, L. S. Chao, L. He, and J. Xing, "Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 365–372. [Online]. Available: http://www.aclweb.org/anthology/W13-2245

[17] N. Q. Luong, B. Lecouteux, and L. Besacier, "LIG system for WMT13 QE task: Investigating the usefulness of features in word confidence estimation for MT," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 396–391.

[18] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting et labeling sequence data," in *Proceedings of ICML-01*, 2001, pp. 282–289.

[19] E. Bicici, "Referential translation machines for quality estimation," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 343–351. [Online]. Available: http://www.aclweb.org/anthology/W13-2242

[20] T. Lavergne, O. Cappé, and F. Yvon, "Practical very large scale crfs," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 504–513.

[21] N. Bach, F. Huang, and Y. Al-Onaizan, "Goodness: A method for measuring machine translation confidence," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, June 19-24 2011, pp. 211–219.

[22] N. Ueffing, K. Macherey, and H. Ney, "Confidence measures for statistical machine translation," in *Proceedings of the MT Summit IX*, New Orleans, LA, September 2003, pp. 394–401.

[23] N. Q. Luong, L. Besacier, and B. Lecouteux, "Word confidence estimation and its integration in sentence quality estimation for machine translation," in *Proceedings of The Fifth International Conference on Knowledge and Systems Engineering (KSE 2013)*, Hanoi, Vietnam, October 17-19 2013.

[24] D. Xiong, M. Zhang, and H. Li, "Error detection for statistical machine translation using linguistic features," in *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, Uppsala, Sweden, July 2010, pp. 604–611.