# Spoken language translation graphs re-decoding using automatic quality assessment

Laurent Besacier, Benjamin Lecouteux, Ngoc-Quang Luong, Ngoc-Tien Le

# SPOKEN LANGUAGE TRANSLATION GRAPHS RE-DECODING USING AUTOMATIC QUALITY ASSESSMENT

*L. Besacier, B. Lecouteux, N.Q. Luong, N.T. Le*

LIG, University of Grenoble, France

## ABSTRACT

This paper investigates how automatic quality assessment of spoken language translation (SLT), also named confidence estimation (CE), can help re-decoding SLT output graphs and improve the overall speech translation performance. Our graph redecoding method can be seen as a second-pass of translation. For this, a robust word confidence estimator for SLT is required. We propose several estimators based on our estimation of transcription (ASR) quality, translation (MT) quality, or both (combined ASR+MT). Using these word confidence measures to re-decode the spoken language translation graph leads to a significant BLEU improvement (more than 2 points) compared to our SLT baseline, for a French-English SLT task. These results could be applied to interactive speech translation or computer-assisted translation of speeches and lectures.

***Index Terms***— Quality estimation, Word confidence estimation (WCE), Spoken Language Translation (SLT), Search graph re-decoding.

## 1. INTRODUCTION

Automatic quality assessment of spoken language translation (SLT), also named confidence estimation (CE), is an important topic because it allows to know if a system produces (or not) user-acceptable outputs. In interactive speech to speech translation, CE helps to judge if a translated turn is uncertain (and ask the speaker to rephrase or repeat). For speech-to-text applications, CE may tell us if output translations are worth being corrected or if they require retranslation from scratch. Building a method which is capable of pointing out the correct parts as well as detecting the errors in a speech translated output is crucial to tackle above issues.

In this paper, we claim that an accurate CE can also help to improve SLT itself through a second-pass N-best list reranking or search graph re-decoding, as it has already been done for text translation by [1] or more recently by [2].

Given signal $x_f$ in the source language, the spoken language translation (SLT) consists in finding the most probable target language sequence $\hat{e} = (e_1, e_2, ..., e_N)$ so that

$$\hat{e} = \underset{e}{\operatorname{argmax}}\{p(e/x_f, f)\} \quad (1)$$

where $f = (f_1, f_2, ..., f_M)$ is the transcription of $x_f$.

Now, if we perform confidence estimation at the "words" level, the problem is called Word-level Confidence Estimation (WCE) and we can represent this information as a sequence $q$ (same length $N$ of $\hat{e}$) where $q = (q_1, q_2, ..., q_N)$ and $q_i \in \{good, bad\}$[1].

Then, integrating automatic quality assessment in our SLT process can be done as following:

$$\hat{e} = \underset{e}{\operatorname{argmax}} \sum_q p(e, q/x_f, f) \quad (2)$$

$$\hat{e} = \underset{e}{\operatorname{argmax}} \sum_q p(q/x_f, f, e) * p(e/x_f, f) \quad (3)$$

$$\hat{e} \approx \underset{e}{\operatorname{argmax}}\{\max_q\{p(q/x_f, f, e) * p(e/x_f, f)\}\} \quad (4)$$

In the product of (4), the SLT component $p(e/x_f, f)$ and the WCE component $p(q/x_f, f, e)$ contribute together to find the best translation output $\hat{e}$.

**Contributions** Following this formalisation, our paper investigates how WCE can help re-decoding SLT output graphs and improve the overall speech translation performance. Our graph redecoding method (discussed in section 2) can be seen as a second-pass of translation. One requirement for this is the availability of a robust confidence estimator based on both ASR and MT features. Our WCE system using such multiple features is discussed in section 3. The experiments described in sections 4 and 5 show how robust WCE helps re-decoding speech translation graphs leading to significant BLEU improvements (more than 2 points). An analysis of the SLT hypotheses with and w/o second pass (based on WCE) is also provided at the end of this paper.

## 2. SLT GRAPH RE-DECODING

### 2.1. Our approach

For practical implementation, we made the following choices for evaluating equation (4):

[1] $q_i$ could be also more than 2 labels, or even scores but this paper only deals with error detection (binary set of labels)

- We want to maximize the product of probabilities over all sequences $q$ of quality tags. This would require applying WCE to the SLT graph but due to technical constraints, our WCE system (detailed later in the paper) can be only applied word sequences. Thus, for each sentence to translate, we tag N-best hypotheses from the SLT system in order to approximate the maximization over $q$.

- For the same reason, the use of WCE labels cannot be integrated directly into the MT decoder so we apply a second pass were labels related to word confidence prediction are integrated into the first-pass search graph (SG) to re-determine the best hypothesis. To do this, our intuition is that all parts of hypotheses corresponding to words labeled as $good$ should be appreciated while those labeled as $bad$ should be weakened.

Consequently, the additional steps (for the second pass) are the following:

- Firstly, apply a WCE classifier on the SLT *N*-best list to assign the quality labels ($good$ or $bad$) along with the confidence probabilities for each word.

- Secondly, for each word in the *N*-best list, update the cost of all SG's hypotheses containing it by adding the updated score to their cost.

- Thirdly, search again on the updated SG for the cheapest-cost hypothesis and thus find the new best translation.

We assume that the decoder generates N best hypotheses $e^N = \{e^1, e^2, ..., e^N\}$ at the end of the first pass. Using the WCE system, we are able to assign the *j-th* word in the hypothesis $e^i$, denoted by $e_{ij}$, with one appropriate quality label, $q_{ij}$. Then, the second pass is carried out by considering every word $e_{ij}$ and its labels $q_{ij}$. Our principal idea is that, if $e_{ij}$ is a *correct* translation, i.e. $q_{ij} = good$, all hypotheses $H_k \in SG$ containing it in the SG should be "rewarded" by reducing their cost. On the contrary, those containing *incorrect* translation will be "penalized". Let $reward(e_{ij})$ and $penalty(e_{ij})$ denote the reward or penalty score of $e_{ij}$, the new transition cost of $H_k$ after being updated is formally defined by:

$$transition'(H_k) = transition(H_k) + \begin{cases} reward(e_{ij}) & \text{if } q_{ij} = good \\ penalty(e_{ij}) & otherwise \end{cases} \quad (5)$$

The update finishes when all words in the *N*-best list have been considered. We then re-compute the new score of complete hypotheses by tracing backward via back-pointers and aggregating the transition cost of all their edges. Essentially, the re-decoding pass reorders SG hypotheses following the rule: the more $good$ words (predicted by WCE system) they contain, the more cost reduction will be made. In this work, the updated scores used are defined as follows:

$$penalty(e_{ij}) = -reward(e_{ij}) = \beta * \frac{score(H_k)}{\#words(H_k)} \quad (6)$$

Where $\#words(H_k)$ is the number of target words in $H_k$, the positive coefficient $\beta$ accounts for the impact level of this score on the hypothesis's final cost and can be optimized during experiments (in this work, due to corpus constraints, we apply a cross-validation procedure where optimization is done on the first half of the test corpus and applied to the second half, and vice-versa). Here, $penalty(e_{ij})$ gets negative sign (since $score(H_k) < 0$) and will be added to the transition cost of all hypotheses containing $e_{ij}$ in case this word is labeled as $bad$; whereas $reward(e_{ij})$ (same value, opposite sign) is used in the other case.

## 3. BUILDING AN EFFICIENT QUALITY ASSESSMENT (WCE) SYSTEM

The WCE component solves the equation:

$$\hat{q} = \underset{q}{\operatorname{argmax}}\{p(q/x_f, f, e)\} \quad (7)$$

where $q = (q_1, q_2, ..., q_N)$ is the sequence of quality labels on the target language. This is a sequence labelling task that can be solved with several machine learning techniques such as Conditional Random Fields (CRF) [3]. However, for that, we need a large amount of training data for which a quadruplet $(x_f, f, e, q)$ is available.

While a corpus containing such a quadruplet was recently made available by [4], it contains only 2.5k utterances which is not sufficient to accurately train the model.

Since this is much easier to obtain data containing either the triplet $(x_f, f, q)$ (automatically transcribed speech with manual references and quality labels infered from word error rate estimation) or the triplet $(f, e, q)$ (automatically translated text with manual post-editions and quality labels infered using tools such as TERpA [5]) we can recast the WCE problem with the following equation :

$$\hat{q} = \underset{q}{\operatorname{argmax}}\{p_{ASR}(q/x_f, f)^{\alpha} * p_{MT}(q/e, f)^{1-\alpha}\} \quad (8)$$

where $\alpha$ is a weight giving more or less importance to $WCE_{ASR}$ (quality assesment on transcription) compared to $WCE_{MT}$ (quality assesment on translation). It is important to note that $p_{ASR}(q/x_f, f)$ corresponds to the quality estimation of the words in the target language based on features calculated on the source language (ASR). For that, what we

do is projecting source quality labels to the target using word-alignment information between $e$ and $f$ sequences. Our two components $WCE_{ASR}$ and $WCE_{MT}$ are more precisley detailed in next subsections (French-English SLT task).

### 3.1. WCE for speech transcription

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 (more details can be found in [4]):

- Acoustic features: word duration (F-dur).
- Graph features (extracted from the ASR word confusion networks): number of alternative (F-alt) paths between two nodes; word posterior probability (F-post).
- Linguistic features (based on probabilities by the language model): word itself (F-word), 3-gram probability (F-3g), back-off level of the targeted word (F-back), as proposed in [6],
- Lexical Features: words Part-Of-Speech (F-POS).

We use a variant of boosting classification algorithm in order to combine features. The used classifier is *bonzaiboost* [7]. 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-POS) previously described. The WCE estimator is trained on a separate french corpus (BREF 120 [8]) which was entirely transcribed automatically to obtain label examples (1M words with labels in total).

### 3.2. WCE for machine translation

We employ the Conditional Random Fields [3] (CRFs) as our machine learning method, with WAPITI toolkit [9], to train the WCE estimator. A separate corpus of 10000 (french-english) MT post-editions (see [10]) is used as training set.

The reason why we use different machine learning technique for confidence estimation in ASR (*boosting*) and in MT (*CRFs*) is due to the fact that these systems were already available before this work. However, our short term goal is to use an unified approach for confidence estimation (for ASR and MT) based only on CRF.

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 [1]: the combinations of the target (source) word and all aligned source (target) words in the window $\pm 2$
- Word posterior probability [11]
- Pseudo-reference (Google Translate): does the word appear in the pseudo reference or not?
- Graph topology [12]: number of alternative paths in the confusion network, 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.
- Lexical features: word Part-Of-Speech (POS); sequence of POS of all its aligned source words; POS bigram (trigram) backward sequences; punctuation; proper name; numerical value.
- Syntactic features: null link [13]; constituent label; depth in the constituent tree after parsing the target hypothesis.
- Semantic features: number of word senses in WordNet (on the target hypothesis).

A very similar feature set was used in our English - Spanish WCE system submitted to WMT 2013 and WMT 2014 Quality Estimation shared task and obtained very good performances [12]. Our experience in participating to the WCE shared task in 2013 and 2014 lead us to the following observation: while feature processing is very important to achieve good performance, it requires to call a set of heterogeneous NLP tools (for lexical, syntactic, semantic analyses). Thus, we recently proposed to unify the feature processing, together with the call of machine learning algorithms, in order to facilitate the design of confidence estimation systems. The open-source toolkit proposed (written in *Python* and made available on *github*[2]) integrates some standard as well as in-house features that have proven useful for WCE (based on our experience in WMT 2013 and 2014). To our knowledge, this is the first toolkit dedicated to word confidence estimation. We also believe that the integrated feature processing of our toolkit could be used for other cross-lingual NLP tasks.

## 4. EXPERIMENTAL SETTINGS

Our dev/test corpus contains 2643 transcribed French speech utterances ($x_f$ and $f_{ref}$ - 5 hours - news domain) translated into English ($e_{ref}$) for which we obtained speech transcription output ($f_{hyp}$) and speech translation output ($e_{hyp}$). Since this corpus is rather small, we apply a cross-validation on it by tuning the re-decoding parameters on half of it while evaluating on the other half, and vice-versa.

To obtain the speech transcripts ($f_{hyp}$), we built an ASR system based on KALDI toolkit [14]. The 3-gram language

[2]http://github.com/besacier/WCE-LIG

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 [15]. In addition, automatic post-processing was needed at the output of the ASR system in order to match requirements of standard input for machine translation (number conversion, recasing, re-punctuating, converting full words back to abbreviations and restoring special characters). With this post-processing, the output of our ASR system, scored against the $f_{ref}$ reference is 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).

To obtain the speech translations ($e_{hyp}$), we used a French-English phrase-based translation system based on *moses* toolkit [16]. This medium-sized system was trained on Europarl and News parallel corpora for a former WMT evaluation shared-task (system more precisely described in [17] - 1.6M parallel sentences and 48M monolingual sentences in target language).

In order to evaluate our WCE system, we obtained a sequence $q$ of quality labels (recall that $q = (q_1, q_2, ..., q_N)$ and $q_i \in \{good, bad\}$) using TERp-A toolkit [5]. Each word or phrase in the hypothesis $e_{hyp}$ is aligned to a word or phrase in the reference ($e_{ref}$) with different types of edit: "I" (insertions), "S" (substitutions), "T" (stem matches), "Y" (synonym matches), "P" (phrasal substitutions) and "E" (exact match). Then, we re-categorize the obtained 6-label set into binary set: the E, T and Y belong to the $good$, whereas the S, P and I belong to the $bad$ category.

Table 1 summarizes MT (translation from manual transcripts $f_{ref}$) and SLT (translation from automatic transcripts $f_{hyp}$) performances obtained on our corpus, as well as the distribution of $good$ and $bad$ labels inferred for both tasks (these labels will be considered as our reference to evaluate WCE later on). Logically, the percentage of (B) labels increases from MT to SLT task in the same conditions.

| **task** | **ASR (WER)** | **MT (BLEU)** | % $good$ | % $bad$ |
|---|---|---|---|---|
| *MT* | 0% | 52.8% | 82.5% | 17.5% |
| *SLT* | 26.6% | 30.6% | 65.5% | 34.5% |

**Table 1**. Baseline MT and SLT performance on 2643 utt.

## 5. EXPERIMENTS ON SLT GRAPH RE-DECODING

### 5.1. Robust estimation of word confidence for a speech translation task

We first report in Table 2 the baseline results obtained by individual WCE system for a single ASR task (second column of the table). Then, we evaluate the performance of 3 WCE systems for the SLT task:

- The first system (SLT sys. / MT feat.) is the one described in section 3.2 and uses only MT features.
- The second system (SLT sys. / ASR feat.) is the one described in section 3.1 and uses only ASR features (so this is predicting SLT output confidence using only ASR confidence features!). Word alignment information between $f_{hyp}$ and $e_{hyp}$ is used to project the WCE scores coming from ASR, to the SLT 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 combined with the MT-based confidence score as shown in equation (8) (we did not tune the $\alpha$ coefficient and set it *a priori* to 0.5).

The results of these 3 systems are given in the last 3 columns of Table 2. They are obtained on the whole test set (all the results are given using a $good/bad$ decision threshold which is a priori set to 0.7). The evaluation metric is the average between the F-measure for $good$ labels and the F-measure for $bad$ labels. From these results, we see that the use of both ASR-based and MT-based confidence scores improve the averaged F-score from 58,25% (MT only features) and 57,20% (ASR only features) to 60,75% (MT+ASR features).

| **task** | **WCE for ASR** | **WCE for SLT** | **WCE for SLT** | **WCE for SLT** |
|---|---|---|---|---|
| feat. type | ASR feat. $p(q/x_f, f)$ | MT feat. $p(q/f, e)$ | ASR feat. $p(q/x_f, f)$ projected to $e$ | MT+ASR feat. $p(q/x_f, f, e)$ |
| *F-mes* | 62,56 % | 58,25% | 57,20% | **60,75%** |

**Table 2**. WCE performance with diff. feat. sets

### 5.2. Re-decoding results

Table 3 compares our 1-pass SLT baseline, to the 2-pass (graph re-decoding) strategy used with three different word confidence estimators: one based on ASR features only; one based on MT features only; and one based on joint MT+ASR features. We see that re-decoding the SLT search graph using WCE improves the translation performance measured with BLEU. The use of joint ASR+MT features for WCE lead to the best performance and the improvement over the 1-pass baseline (more than 2 BLEU points) is significant ($p \in [0.00; 0.01]$ evaluated according to [18]). Comparing the 2-pass results, we observe that ASR+MT features clearly overpass MT features (BLEU is 32.82% instead of 31.89%) which may seem surprising because results of table 2 did not show a huge difference between both WCE methods. A first explanation may be related to the fact that for re-decoding, the confidence estimator is not applied on 2643 sentences only, but on 2643 * N sentences (in this case N = 100 best hypotheses) and differences in performance between MT and

ASR+MT may be more important in this case. Another explanation may be also that even a small improvement in error detection (words whose label is $bad$) can lead to a significant gain of BLEU score. Finally, the use of ASR features only improves the performance compared to a single-pass system (31.12% instead of 30.60%) but this is the weakest improvement observed.

| system | baseline | redecoding | redecoding | redecoding |
|---|---|---|---|---|
| **WCE feat.** | none | ASR $p(q/x_f, f)$ | MT $p(q/f, e)$ | SLT $p(q/x_f, f, e)$ |
| *Perf.* | 30.60% | 31.12% | 31.89% | **32.82%** |

**Table 3**. SLT perf. (BLEU) after 2d pass (2643 utt.)

### 5.3. Analysis of SLT hypotheses

| | |
|---|---|
| **example 1** $f_{ref}$ | une démobilisation des employés peut déboucher sur une démoralisation **mortifère** |
| $f_{hyp}$ | une démobilisation des employés peut déboucher sur une démoralisation **mort y faire** |
| $e_{hyp}$ baseline | a **demobilisation employees** can lead to a **penalty demoralisation** |
| $e_{hyp}$ with redecoding | a **demobilisation of employees** can lead to a **demoralization death** |
| $e_{ref}$ | **demobilization of employees** can lead to a **deadly demoralization** |
| **example 2** $f_{ref}$ | celui-ci a indiqué que l'intervention **s'**était parfaitement bien **déroulée** et que les examens post-**opératoires** étaient normaux |
| $f_{hyp}$ | celui-ci a indiqué que l' intervention **c'**était parfaitement bien **déroulés** , et que les examens post **opératoire** étaient normaux . |
| $e_{hyp}$ baseline | **it** has indicated that the speech **that was well** conducted , and that the tests were **normal post route** |
| $e_{hyp}$ with redecoding | **he** indicated that the intervention **is very well done** , and that the tests **after operating were normal** |
| $e_{ref}$ | **he** indicated that the operation **went perfectly well** and the **post-operative tests were normal** |
| **example 3** $f_{ref}$ | general motors repousse jusqu'en janvier le plan pour **opel** |
| $f_{hyp}$ | general motors repousse jusqu' en janvier le plan pour **open** |
| $e_{hyp}$ baseline | general motors postponed until january **the plan to open** |
| $e_{hyp}$ with redecoding | general motors puts until january **terms to open** |
| $e_{ref}$ | general motors postponed until january **the plan for opel** |

**Table 4**. Exemples of French SLT hyp with and w/o redecoding

Examples of speech translation hypotheses (SLT) obtained with or without graph re-decoding are given in table 4 (without trying to analyze fine differences between MT and ASR+MT estimators - so, line indicating *with re-decoding* corresponds to the use of one or the other estimator).

Example 1 illustrates a first case where re-decoding allows slightly improving the translation hypothesis. Analysis of the labels from the confidence estimator indicates that the words *a* (start of sentence) and *penalty* were labeled as $bad$ here. Thus, a better hypothesis arised from the second pass, although the transcription error could not be recovered (since only the 1-best ASR hypothesis is translated so far - not the full ASR graph).

In example 2, the confidence estimator labeled as $bad$ the following word sequences: *it has*, *speech that was* and *post route*. Better translation hypothesis is found after re-decoding (correct pronoun, better quality at the end of sentence).

Finally, example 3 shows a case where, this time, the end of the first pass translation deteriorated after re-decoding. Analysis of confidence estimator output shows that the phrase *to open* was (correctly) labeled as $bad$, but the re-decoding gave rise to an even worse hypothesis. This last example illustrates some limitations of our current approach, which in this case would - in any event - have been unable to recover the named entity *opel* which was not present in the translation graph.

## 6. RELATED WORK

Several previous works tried to propose effective confidence measures in order to detect errors on ASR outputs. Confidence measures are introduced for Out-Of-Vocabulary (OOV) detection by [19]. [20] extends the previous work and introduces the use of word posterior probability (WPP) as a confidence measure for speech recognition. Posterior probability of a word is most of the time computed using the hypothesis word graph [21]. Also, recent approaches [22] for confidence measure estimation use side-information extracted from the recognizer: normalized likelihoods (WPP), number of competitors at the end of a word (hypothesis density), decoding process behavior, linguistic features, acoustic (acoustic stability, duration features) and semantic features. In parallel, the Workshop on Machine Translation (WMT) introduced in 2013 a WCE task for machine translation. [23] [24] employed the Conditional Random Fields (CRF) [3] model as their Machine Learning method to address the problem as a sequence labelling task. Meanwhile, [25] extended the global learning model by dynamic training with adaptive weight updates in the perceptron training algorithm. As far as prediction indicators are concerned, [25] proposed seven word feature types and found among them the "common cover links" (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. [23] focused only on various n-gram combinations of target words. Inheriting most of previously-recognized features, [24] 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. To our knowledge, the first attempt to design WCE for speech translation, using joint ASR and MT features, is the work published very recently at IWSLT 2014 by [4].

Concerning second-pass MT, several works dealt with automatic post-edition [26] or N-best re-ranking [27, 1, 28].

[29] compared N-best re-ranking with graph re-decoding for statistical machine translation and have shown that graph re-decoding is more efficient. This is why this paper focus on graphs instead of N-best lists.

If we focus on graph re-decoding, [30] proposes a 2-pass translation system which uses, in the second pass, sentence length and word sequence probability parameters. Experiments for Mandarin-English MT (NIST task) show a significant improvement in performance. Moreover, [31] proposes to re-decode the MT graph using minimization of Bayes risk (MBR). The experimental results on Arabic-English, Mandarin-English and English-Mandarin show that the approach outperforms a N-best re-ranking baseline (also using MBR criterion). Finally, [32] uses a N-gram language model (with longer history) to re-decode a translation graph obtained using probabilistic context free grammar (PCFG) MT. More recently, [33] integrated a large-scale neural language model into a machine translation system both by reranking N-best lists and by direct integration into the decoder. All these works were applied to MT graphs re-decoding only (no actual speech translation).

In addition to being applied to spoken language (SLT), our approach is different of the above mentionned since we are using, for second-pass decoding, external information features gathered via a robust confidence estimator (based on combined ASR and MT features). However, it present some similarities with [30] but using a much bigger number of features (for assessing word confidence).

## 7. CONCLUSION

We have proposed a formalisation of the use of quality assessment in speech translation. Moreover, experiments have shown that WCE (word confidence estimation) labels can be used to successfully re-decode the speech translation graphs and significantly improve speech translation performance. The use of combined MT and ASR features for robust word confidence estimation in SLT lead to the best performance in second-pass re-decoding. Some perspectives of this work are the following: train a unique WCE system for SLT (evaluating $p(q/x_f, f, e)$ as in equation (7)) using joint ASR+MT features and see if more SLT errors can be accurately detected, re-decoding speech translation graph obtained after translating the full ASR lattice (so far, a SLT graph is obtained by translating only the 1-best of the ASR), use our approach for real interactive speech translation scenarios such as news or lectures subtitling.

## 8. REFERENCES


[1] Nguyen Bach, Fei Huang, and Yaser 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.

[2] Ngoc-Quang Luong, Laurent Besacier, and Benjamin Lecouteux, "Word Confidence Estimation for SMT N-best List Re-ranking," in *Proceedings of the Workshop on Humans and Computer-assisted Translation (HaCaT) during EACL*, Gothenburg, Suède, 2014.

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

[4] Laurent Besacier, Benjamin Lecouteux, Ngoc Quang Luong, Kaing Hour, and Marwa Hadjsalah, "Word confidence estimation for speech translation," in *Proceedings of The International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, USA, December 2014.

[5] Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz, "Terp system description," in *MetricsMATR workshop at AMTA*, 2008.

[6] Julien Fayolle, Fabienne Moreau, Christian Raymond, Guillaume Gravier, and Patrick Gros, "Crf-based combination of contextual features to improve a posteriori word-level confidence measures.," in *Interspeech*, 2010.

[7] Antoine Laurent, Nathalie Camelin, and Christian Raymond, "Boosting bonsai trees for efficient features combination : application to speaker role identification," in *Interspeech*, 2014.

[8] Lori F Lamel, Jean-Luc Gauvain, Mazcine Eskénazi, et al., "Bref, a large vocabulary spoken corpus for french1," *training*, vol. 22, no. 28, pp. 50, 1991.

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

[10] 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.

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

[12] Ngoc Quang Luong, Laurent Besacier, and Benjamin 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.

[13] Deyi Xiong, Min Zhang, and Haizhou 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.

[14] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[15] Sylvain Galliano, Edouard Geoffrois, Guillaume Gravier, Jean-Francois Bonastre, Djamel Mostefa, and Khalid 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.

[16] Philipp Koehn, 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, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, June 2007, pp. 177–180.

[17] 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)*, ACL Workshop, Ed., Uppsala, Sweden, 11-17 July 2010.

[18] Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith, "Better hypothesis testing for statistical machine translation: Controlling for optimizer instability," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, Association for Computational Linguistics, Ed., Portland, Oregon, 2011, pp. 176–181.

[19] Ayman Asadi, Richard Schwartz, and John 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.

[20] Sheryl 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.

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

[22] Benjamin Lecouteux, Georges Linarès, and Benoit Favre, "Combined low level and high level features for out-of-vocabulary word detection," *INTERSPEECH*, 2009.

[23] Aaron Li-Feng Han, Yi Lu, Derek F. Wong, Lidia S. Chao, Liangye He, and Junwen 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, August 2013, pp. 365–372, Association for Computational Linguistics.

[24] Ngoc Quang Luong, Benjamin Lecouteux, and Laurent 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, August 2013, pp. 396–391, Association for Computational Linguistics.

[25] Ergun Bicici, "Referential translation machines for quality estimation," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August 2013, pp. 343–351, Association for Computational Linguistics.

[26] Kristen Parton, Nizar Habash, Kathleen McKeown, Gonzalo Iglesias, and Adrià de Gispert, "Can automatic post-editing make mt more meaningful?," in *Proceedings of the 16th EAMT*, Trento, Italy, 28-30 May 2012, pp. 111–118.

[27] Kevin Duh and Katrin Kirchhoff, "Beyond log-linear models: Boosted minimum error rate training for n-best re-ranking," in *Proc. of ACL, Short Papers*, 2008.

[28] Ying Zhang, Almut Silja Hildebrand, and Stephan Vogel, "Distributed language modeling for n-best list re-ranking," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, Sydney, July 2006, pp. 216–223.

[29] Ngoc Quang Luong, Laurent Besacier, and Benjamin Lecouteux, "An Efficient Two-Pass Decoder for SMT Using Word Confidence Estimation," in *European Association for Machine Translation (EAMT)*, Dubrovnik, Croatie, jun 2014.

[30] Richard Zens and Hermann Ney, "N-gram posterior probabilities for statistical machine translation," in *Workshop on Statistical Machine Translation - StatMT*, Stroudsburg, PA, USA, 2006.

[31] Roy Tromble, Shankar Kumar, Franz Josef Och, and Wolfgang Macherey, "Lattice minimum bayes risk decoding for statistical machine translation.," in *Lattice minimum bayesrisk Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2008, pp. 620–629.

[32] Ashish Venugopal, Andreas Zollmann, and Stephan Vogel, "An efficient two-pass approach to synchronous-cfg driven statistical mt," in *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, April 2007.

[33] Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang, "Decoding with large-scale neural language models improves translation," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.