Joint ASR and MT Features for Quality Estimation in Spoken Language Translation

Ngoc-Tien Le, Benjamin Lecouteux and Laurent Besacier

GETALP – LIG, University of Grenoble Alpes, France

firstname.lastname@imag.fr

Abstract

This paper aims to unravel the automatic quality assessment for spoken language translation (SLT). More precisely, we propose several effective estimators based on our estimation of transcription (ASR) quality, translation (MT) quality, or both (combined and joint features using ASR and MT information). Our experiments provide an important opportunity to advance the understanding of the prediction quality of words in a SLT output that were revealed by MT and ASR features. These results could be applied to interactive speech translation or computer-assisted translation of speeches and lectures. For reproducible experiments, the code allowing to call our WCE-LIG application and the corpora used are made available to the research community.

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. Moreover, an accurate CE can also help to improve SLT itself through a second-pass N-best list re-ranking or search graph re-decoding, as it has already been done for text translation in [1] and [2], or for speech translation in [3]. Consequently, 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.

Outline The outline of this paper goes simply as follows: section 2 reviews the state-of-the-art on confidence estimation for ASR and MT. Our word confidence estimation (WCE) system using multiple features is then described in section 3. The experimental setup (notably our specific WCE corpus) is presented in section 4 while section 5 evaluates our joint WCE system and finally, section 6 concludes this work and gives some perspectives.

2. Related Work on Confidence Estimation for ASR and MT

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 [4]. [5] 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 [6]. Also, more recent approaches [7] 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, linguistic features, acoustic features (acoustic stability, duration features) and semantic features.

In parallel, the Workshop on Machine Translation (WMT) introduced in 2013 a WCE task for Machine Translation. [8] [9] employed the Conditional Random Fields (CRF) [10] model as their machine learning method to address the problem as a sequence labelling task. Meanwhile, [11] extended their initial proposition by dynamic training with adaptive weight updates in their neural network classifier. As far as prediction indicators are concerned, [11] 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. [8] focused only on various n-gram combinations of target words. Inheriting most of previously-recognized features, [9] 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. The estimation of the confidence score uses mainly classifiers like Conditional Random Fields [8, 12], Support Vector Machines [13] or Perceptron [11]. Some investigations were also conducted to determine which features seem to be the most relevant. [13] proposed to filter features using a forward-backward algorithm to discard linearly correlated features. Using Boosting as learning algorithm, [14] was able to take advantage of the most significant features.

Finally, several toolkits for WCE were recently propo-

sed: *TranscRater* for ASR [15] [1], Marmot for MT [2] as well as WCE toolkit [16] [3] that will be used to extract MT features in the experiments of this paper.

To our knowledge, the first attempt to design WCE for speech translation, using both ASR and MT features, is our own work [17, 3] which is further extended in this paper submission.

3. Building an Efficient Quality Assessment (WCE) System

The WCE component solves the equation:

$$\hat{q} = \underset{q}{\operatorname{argmax}}\{p_{SLT}(q|x_f, f, e)\} \quad (1)$$

where x_f is the given signal in the source language, spoken language translation (SLT) consists in finding the most probable target language sequence $\hat{e} = (e_1, e_2, ..., e_N)$; $f = (f_1, f_2, ..., f_M)$ is the transcription of x_f; $q = (q_1, q_2, ..., q_N)$ is the sequence of quality labels on the target language and $q_i \in \{good, bad\}$ [4]. This is a sequence labelling task that can be solved with several machine learning techniques such as Conditional Random Fields (CRF) [10]. However, for that, we need a large amount of training data for which a quadruplet (x_f, f, e, q) is available. In this work, we will use a corpus extended from [17] which contains 6.7k utterances. We will investigate if this amount of data is enough to evaluate and test a joint model $p_{SLT}(q|x_f, f, e)$.

As it 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 TERp-A [18]) we can also recast the WCE problem with the following equation:

$$\hat{q} = \underset{q}{\operatorname{argmax}}\{p_{ASR}(q|x_f, f)^{\alpha} * p_{MT}(q|e, f)^{1-\alpha}\} \quad (2)$$

where α is a weight giving more or less importance to WCE_{ASR} (quality assessment on transcription) compared to WCE_{MT} (quality assessment 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 scores to the target using word-alignment information between e and f sequences. This alternative approach (equation 2) will be also evaluated in this work even if it corresponds to a different optimization problem than equation 1. In particular, the choice of α is only set a priori in our experiments to 0.5 which is probably not the best option.

In both approaches – *joint* ($p_{SLT}(q|x_f, f, e)$) and *combined* ($p_{ASR}(q|x_f, f)$ + $p_{MT}(q|e, f)$) – some features need to be extracted from ASR and MT modules. They are more precisely detailed in next subsections.

3.1. WCE Features for Speech Transcription (ASR)

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 [17]):

- 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**), log probability (**F-log**), back-off level of the word (**F-back**), as proposed in [19],
- Lexical Features: Part-Of-Speech (POS) of the word (**F-POS**),
- Context Features: Part-Of-Speech tags in the neighborhood of a given word (**F-context**).

For each word in the ASR hypothesis, we estimate the 9 features (F-Word ; F-3g ; F-back ; F-log ; F-alt ; F-post ; F-dur ; F-POS ; F-context) previously described.

In a preliminary experiment, we will evaluate these features for quality assessment in ASR only (WCE_{ASR} task). Two different classifiers will be used: a variant of boosting classification algorithm called *bonzaiboost* [20] (implementing the boosting algorithm *Adaboost.MH* over deeper trees) and the Conditional Random Fields [10].

3.2. WCE Features for Machine Translation (MT)

A number of knowledge sources are employed for extracting features, in a total of 24 major feature types, see Table 1.

It is important to note that we extract features regarding *tokens* in the Machine Translation (MT) hypothesis sentence. In other words, one feature is extracted for each token in the MT output. So, in the Table 1, *target* refers to the feature coming from the MT hypothesis and *source* refers to a feature extracted from the source word aligned to the considered target word. More details on some of these features are given in the next subsections.

3.2.1. Internal Features

These features are given by the Machine Translation system, which outputs additional data like N-best list.

1. https://github.com/hlt-mt/TranscRater
2. https://github.com/qe-team/marmot
3. https://github.com/besacier/WCE-LIG
4. q_i could be also more than 2 labels, or even scores but this paper only deals with error detection (binary set of labels).

1 Proper Name	10 Stop Word	19 WPP max
2 Unknown Stem	11 Word context Alignments	20 Nodes
3 Num. of Word Occ.	12 POS context Alignments	21 Constituent Label
4 Num. of Stem Occ.	13 Stem context Alignments	22 Distance To Root
5 Polysemy Count – Target	14 Longest Target N-gram Length	23 Numeric
6 Backoff Behaviour – Target	15 Longest Source N-gram Length	24 Punctuation
7 Alignment Features	16 WPP Exact	
8 Occur in Google Translate	17 WPP Any	
9 Occur in Bing Translator	18 WPP min	

Table 1 – List of MT features extracted.

Word Posterior Probability (WPP) and **Nodes** features are extracted from a confusion network, which comes from the output of the Machine Translation N-best list. **WPP Exact** is the WPP value for each word concerned at the exact same position in the graph. **WPP Any** extracts the same information at any position in the graph. **WPP Min** gives the smallest WPP value concerned by the transition and **WPP Max** its maximum.

3.2.2. External Features

Below is the list of the external features used:

- **Proper Name**: indicates if a word is a proper name (same binary features are extracted to know if a token is **Numerical**, **Punctuation** or **Stop Word**).
- **Unknown Stem**: informs whether the stem of the considered word is known or not.
- **Number of Word/Stem Occurrences**: counts the occurrences of a word/stem in the sentence.
- **Alignment context features**: these features (#11-13 in Table 1) are based on collocations and proposed by [1].
- **Longest Target (or Source) N-gram Length**: we seek to get the length $(n + 1)$ of the longest left sequence (w_{i-n}) concerned by the current word (w_i) and known by the language model (LM) concerned (source and target sides). We also extract a redundant feature called **Backoff Behavior Target**.
- The target word's constituent label (**Constituent Label**) and its depth in the constituent tree (**Distance to Root**) are extracted using a syntactic parser.
- **Target Polysemy Count**: we extract the polysemy count, which is the number of meanings of a word in a given language.
- **Occurences in Google Translate** and **Occurences in Bing Translator**: in the translation hypothesis, we (optionally) test the presence of the target word in online translations given respectively by *Google Translate* and *Bing Translator*[5].

5. Using this kind of feature is controversial, however we observed that such features are available in general use case scenarios, so we decided to include them in our experiments. Contrastive results without these 2 features will be also given later on.

A very similar feature set was used for a simple WCE_{MT} task (English - Spanish MT, WMT 2013, 2014 quality estimation shared task) and obtained very good performances [21].

In this paper, we will use only Conditional Random Fields [10] (CRFs) as our machine learning method, with WAPITI toolkit [22], to train our WCE estimator based on MT and ASR features.

4. Experimental Setup

4.1. Dataset

The *dev* set and *tst* set of this corpus were recorded by french native speakers. Each sentence was uttered by 3 speakers, leading to 2643 and 4050 speech recordings for *dev* set and *tst* set, respectively. For each speech utterance, a quintuplet containing: ASR output (f_{hyp}), verbatim transcript (f_{ref}), English text translation output ($e_{hyp_{mt}}$), speech translation output ($e_{hyp_{slt}}$) and post-edition of translation (e_{ref}), was made available. This corpus is available on a *github* repository[6]. The total length of the *dev* and *tst* speech corpus obtained are 16h52, since some utterances were pretty long.

4.2. ASR Systems

To obtain the speech transcripts (f_{hyp}), we built a French ASR system based on KALDI toolkit [23]. Acoustic models are trained using several corpora (ESTER, REPERE, ETAPE and BREF120) representing more than 600 hours of french transcribed speech.

We propose to use two 3-gram language models trained on French ESTER corpus [24] as well as on French Gigaword (vocabulary size are respectively 62k and 95k). The ASR systems LM weight parameters are tuned through WER on the *dev* corpus.

Table 2 presents the performances obtained by two above ASR systems.

These WER may appear as rather high according to the task (transcribing read news). A deeper analysis shows that these news contain a lot of foreign named entities, especially in our *dev* set. This part of the data is extracted

6. `https://github.com/besacier/WCE-SLT-LIG/`

Task	*dev* set	*tst* set
ASR1	21.86%	17.37%
ASR2	16.90%	12.50%

Table 2 – ASR performance (WER) on our *dev* and *test* set for the two different ASR systems.

from French medias dealing with european economy in EU. This could also explain why the scores are significantly different between *dev* and *test* sets. In addition, automatic post-processing is applied to ASR output in order to match requirements of standard input for Machine Translation.

4.3. SMT System

We used *moses* phrase-based translation toolkit [25] to translate French ASR into English (e_{hyp}). This medium-size system was trained using a subset of data provided for IWSLT 2012 evaluation [26]: Europarl, Ted and News-Commentary corpora. The total amount is about 60M words. We used an adapted target language model trained on specific data (News Crawled corpora) similar to our evaluation corpus (see [27]). This standard SMT system will be used in all experiments reported in this paper.

4.4. Obtaining Quality Assessment Labels for SLT

After building an ASR system, we have a new element of our desired quintuplet: the ASR output f_{hyp}. It is the noisy version of our already available verbatim transcripts called f_{ref}. This ASR output (f_{hyp}) is then translated by the exact same SMT system [27] already mentioned in subsection 4.3. This new output translation is called $e_{hyp_{slt}}$ and it is a degraded version of $e_{hyp_{mt}}$ (translation of f_{ref}).

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 e_{ref}), to infer the quality (G, B) labels of our speech translation output $e_{hyp_{slt}}$. The word label setting for WCE is done using TERp-A toolkit [18] between $e_{hyp_{slt}}$ and e_{ref}. This assumption, and 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 and Table 5) where it is shown that there is not a huge difference between the MT and SLT performance (evaluated with BLEU).

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 [7] contain also a quintuplet with ASR output, verbatim transcript, MT output, SLT output and target translation. But there are 2 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 *good/bad* labels extracted from this would be reliable for WCE training and testing) ; secondly, in our corpus, each sentence is uttered by 3 different speakers which introduces speaker variability in the database and allows us to deal with different ASR outputs for a single source sentence.

7. http://www-lium.univ-lemans.fr/fr/content/corpus-ted-lium

4.5. Final Corpus Statistics

The final corpus obtained is summarized in Table 3, 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:

- **ASR**: extract *good/bad* labels by calculating WER between f_{hyp} and f_{ref},
- **MT**: extract *good/bad* labels by calculating TERp-A between $e_{hyp_{mt}}$ and e_{ref},
- **SLT**: extract *good/bad* labels by calculating TERp-A between $e_{hyp_{slt}}$ and e_{ref}.

Table 4 and Table 5 summarize baseline ASR, MT and SLT performances obtained on our corpora, 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.

5. Experiments on WCE for SLT

5.1. SLT Quality Assessment Using only MT or ASR Features

We first report in Table 6 the baseline WCE results obtained using MT or ASR features separately. In short, we evaluate the performance of 4 WCE systems for different tasks:

- The first and second systems (WCE for ASR / ASR feat.) use ASR features described in section 3.1 with two different classifiers (CRF or Boosting).
- The third system (WCE for SLT / MT feat.) uses only MT features described in section 3.2 with CRF classifier.
- The fourth system (WCE for SLT / ASR feat.) uses only ASR features described in section 3.1 with CRF classifier (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.

In all experiments reported in this paper, we evaluate the performance of our classifiers by using the average between the F-measure for *good* labels and the F-measure for *bad* labels that are calculated by the common evaluation metrics: Precision, Recall and F-measure for *good/bad* labels. Since two ASR systems are available, *F-mes1* is obtained for SLT based on $ASR1$ whereas *F-mes2* is obtained for SLT based on $ASR2$. For the results of Table 6, the classifier is evaluated on the *tst* part of our corpus and trained on the *dev* part.

Data	# *dev* utt	# *test* utt	# *dev* words	# *test* words	method to obtain WCE labels
f_{ref}	881	1350	21 988	36 404	
f_{hyp1}	881*3	1350*3	66 435	108 332	wer(f_{hyp1}, f_{ref})
f_{hyp2}	881*3	1350*3	66 834	108 598	wer(f_{hyp2}, f_{ref})
$e_{hyp_{mt}}$	881	1350	22 340	35 213	terpa($e_{hyp_{mt}}$, e_{ref})
$e_{hyp_{slt1}}$	881*3	1350*3	61 787	97 977	terpa($e_{hyp_{slt1}}$, e_{ref})
$e_{hyp_{slt2}}$	881*3	1350*3	62 213	97 804	terpa($e_{hyp_{slt2}}$, e_{ref})
e_{ref}	881	1350	22 342	34 880	

Table 3 – Overview of our post-edition corpus for SLT.

Task	ASR (WER)	MT (BLEU)	% G (good)	% B (bad)
MT	0%	49.13%	76.93%	23.07%
SLT (ASR1)	21.86%	26.73%	62.03%	37.97%
SLT (ASR2)	16.90%	28.89%	63.87%	36.13%

Table 4 – MT and SLT performances on our *dev* set.

Concerning WCE for ASR, we observe that F-measure decreases when ASR WER is lower (*F-mes2*<*F-mes1* while $WER_{ASR2} < WER_{ASR1}$). So quality assessment in ASR seems to become harder as the ASR system improves. This could be due to the fact that the ASR1 errors recovered by bigger LM in ASR2 system were easier to detect. The effect of the classifier (CRF or Boosting) is not conclusive since CRF is better for *F-mes1* and worse for *F-mes2*.

As can be seen from the results of WCE for SLT, we can see that F-measure is better using MT features rather than ASR features (quality assessment for SLT more dependent of MT features than ASR features). Again, F-measure decreases when ASR WER is lower (*F-mes2*<*F-mes1* while $WER_{ASR2} < WER_{ASR1}$).

In the next subsection, we try to see if the use of both MT and ASR features improves quality assessment for SLT.

5.2. SLT Quality Assessment Using both MT and ASR Features

We now report in Table 6 WCE for SLT results obtained using both MT and ASR features. More precisely we evaluate two different approaches (*combination* and *joint*):

- The first system (WCE for SLT / MT+ASR feat.) combines the output of two separate classifiers based on ASR and MT features. In this approach, 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 2* (we did not tune the α coefficient and set it *a priori* to 0.5).
- The second system (joint feat.) trains a single WCE system for SLT (evaluating $p(q|x_f, f, e)$ as in *equation 1* using joint ASR features and MT features. All ASR features are projected to the target words using automatic word alignments. However, a problem occurs when a target word does not have any source word aligned to it. In this case, we decide to duplicate the ASR features of its previous target word. Another problem occurs when a target word is aligned to more than one source word. In that case, there are several strategies to infer the 9 ASR features: the average value for F-post, F-log, F-back and the maximum value for F-3g, F-alt, F-dur. For the other features, we generate the values of the first source word aligned to it.

The results of Table 6 show that joint ASR and MT features do not improve WCE performance: *F-mes1* and *F-mes2* are slightly worse than those of Table 6 (WCE for SLT / MT features only). We also observe that simple combination (MT+ASR) degrades the WCE performance. This latter observation may be due to different behaviors of WCE_{MT} and WCE_{ASR} classifiers which makes the weighted combination ineffective. Moreover, the disappointing performance of our joint classifier may be due to an insufficient training set (only 2643 utterances in *dev* !). Finally, removing *OccurInGoogleTranslate* and *OccurInBingTranslate* features for *Joint* lowered *F-mes* between 1% and 1.5%.

These observations lead us to investigate the behaviour of our WCE approaches for a large range of $good/bad$ decision threshold.

While the previous tables provided WCE performance for a single point of interest ($good/bad$ decision threshold set to 0.5), the curves of figures 1 show the full picture of our WCE systems (for SLT) using speech transcriptions systems $ASR1$ and $ASR2$, respectively. We observe that the classifier based on ASR features has a very different behaviour than the classifier based on MT features which explains why their simple combination (MT+ASR) does not work very well for the default decision threshold (0.5). However, for threshold above 0.75, the use of both ASR and MT features is slightly beneficial. This is interesting because higher thre-

Task	**ASR (WER)**	**MT (BLEU)**	**% G (good)**	**% B (bad)**
MT	0%	57.87%	81.58%	18.42%
SLT ($ASR1$)	17.37%	30.89%	61.12%	38.88%
SLT ($ASR2$)	12.50%	33.14%	62.77%	37.23%

Table 5 – MT and SLT performances on our *tst* set.

task feat. type	**WCE for ASR** ASR feat. $p(q\|x_f, f)$ (CRFs)	**WCE for ASR** ASR feat. $p(q\|x_f, f)$ (Boosting)	**WCE for SLT** MT feat. $p(q\|f, e)$	**WCE for SLT** ASR feat. $p_{ASR}(q\|x_f, f)$ projected to e	**WCE for SLT** MT+ASR feat. $p_{ASR}(q\|x_f, f)^{\alpha}$ $*p_{MT}(q\|e, f)^{1-\alpha}$	**WCE for SLT** Joint feat. $p(q\|x_f, f, e)$
F-mes1	68.71%	64.27%	60.55%*	49.67%	52.99%	60.29%**
F-mes2	59.83%	62.61%	59.83%*	44.56%	48.46%	59.23%**

Table 6 – WCE performance with different feature sets for *tst* set (training is made on *dev* set) - After removing *OccurInGoogleTranslate* and *OccurInBingTranslate* features: * for MT feat, lead to 59.40% and 58.11% for *F-mes1* and *F-mes2* respectively ; ** for *joint feature*, lead to 59.14% and 57.75% for *F-mes1* and *F-mes2* respectively.

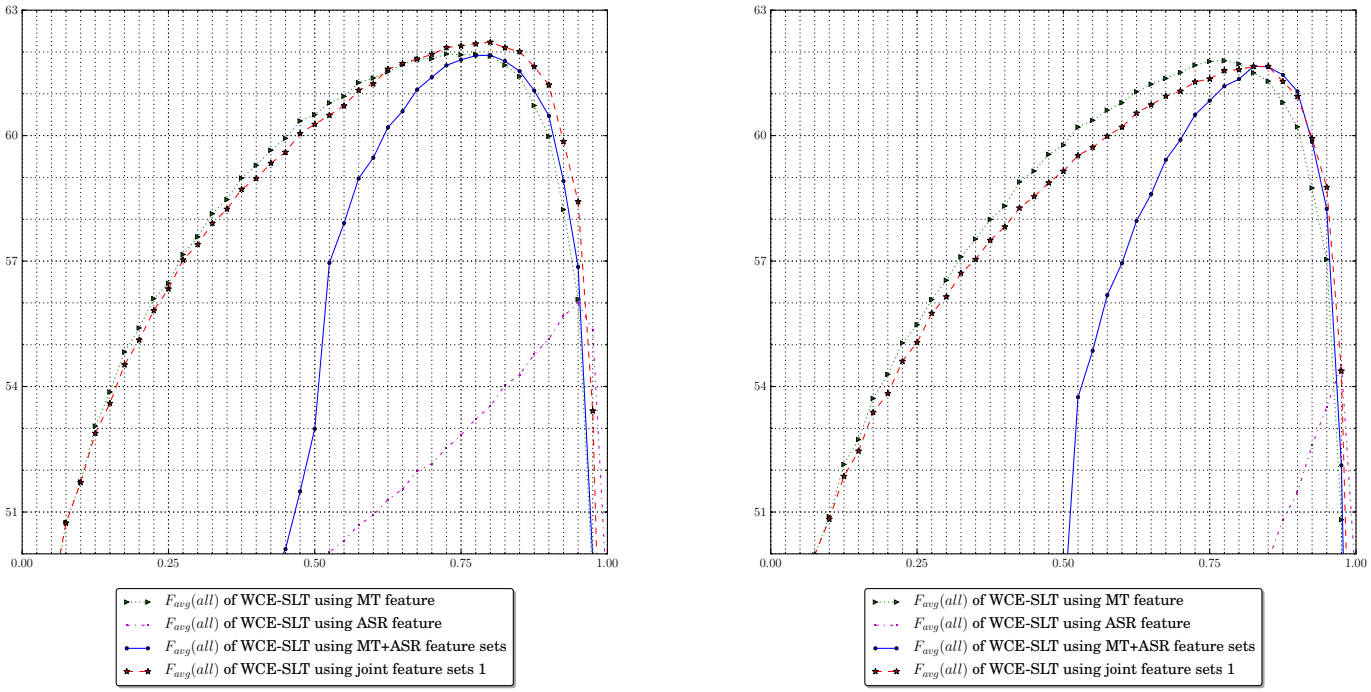

Figure 1 – Evolution of system performance (y-axis - *F-mes1* - ASR1 and *F-mes2* - ASR2) for *tst* corpus (4050 utt) along decision threshold variation (x-axis) - training is made on *dev* corpus (2643 utt).

sholds improves the F-measure on *bad* labels (so improves error detection). Both curves are similar whatever the ASR system used. These results suggest that with enough development data for appropriate threshold tuning (which we do not have for this very new task), the use of both ASR and MT features should improve error detection in speech translation (blue and red curves are above the green curve for higher decision threshold [8]).

8. Corresponding to optimization of the F-measure on *bad* labels (errors).

6. Conclusion

6.1. Main Contributions

In this paper, we introduced a new quality assessment task: word confidence estimation (WCE) for spoken language translation (SLT). A specific corpus, distributed to the research community [9] was built for this purpose. We formalized WCE for SLT and proposed several approaches based on several types of features: Machine Translation (MT) based features, automatic speech recognition (ASR) based features, as well as combined or joint features using ASR and MT information. The proposition of a unique *joint* classifier based on different feature types (ASR and MT features) could allow to operate feature selection in the future and analyze which features (from ASR or MT) are the most efficient for quality assessment in speech translation. Our experiments have shown that MT features remain the most influential while ASR features can bring interesting complementary information. In all our experiments, we systematically evaluated with two ASR systems that have different performance in order to analyze the behavior of our quality assessment algorithms at different levels of word error rate (WER). This allowed us to observe that WCE performance decreases as ASR system improves. For reproducible research, most features [10] and algorithms used in this paper are available through our toolkit called WCE-LIG. This package is made available on a *GitHub* repository [11] under the licence GPL V3. We hope that the availability of our corpus and toolkit could lead, in a near future, to a new shared task dedicated to quality estimation for speech translation. Such a shared task could be proposed in avenues such as IWSLT (International Workshop on Spoken Language Translation) or WMT (Workshop on Machine Translation) for instance.

6.2. Other Perspectives

In addition to re-decode SLT graphs, our quality assessment system can be used in interactive speech translation scenarios such as news or lectures subtitling, to improve human translator productivity by giving him/her feedback on automatic transcription and translation quality. Another application would be the adaptation of our WCE system to interactive speech-to-speech translation scenarios where feedback on transcription and translation modules is needed to improve communication. On these latter subjects, it would also be nice to move from a binary ($good$ or bad labels) to a 3-class decision problem (*good, asr-error, mt-error*). The outcome material of this paper (corpus, toolkit) can be definitely used to address such a new problem.

9. https://github.com/besacier/WCE-SLT-LIG
10. MT features already available, ASR features available soon
11. https://github.com/besacier/WCE-LIG

7. References

[1] 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.

[2] N.-Q. Luong, L. Besacier, and B. 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. [Online]. Available: http://hal.inria.fr/hal-00953719

[3] L. Besacier, B. Lecouteux, N.-Q. Luong, and N.-T. Le, "Spoken language translation graphs re-decoding using automatic quality assessment," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scotsdale, Arizona, United States, Dec. 2015. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01289158

[4] 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.

[5] 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.

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

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

[8] 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

[9] 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.

[10] 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.

[11] 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

[12] N.-Q. Luong, L. Besacier, and B. Lecouteux, "LIG System for Word Level QE task at WMT14," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland USA, June 2014, pp. 335–341.

[13] D. Langlois, S. Raybaud, and K. Smaïli, "Loria system for the wmt12 quality estimation shared task," in *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Baltimore, Maryland USA, June 2012, pp. 114–119.

[14] N.-Q. Luong, L. Besacier, and B. Lecouteux, "Towards accurate predictors of word quality for machine translation: Lessons learned on french - english and english - spanish systems," *Data and Knowledge Engineering*, p. 11, Apr. 2015.

[15] S. Jalalvand, M. Negri, M. Turchi, J. G. C. de Souza, F. Daniele, and M. R. H. Qwaider, "Transcrater: a tool for automatic speech recognition quality estimation," in *Proceedings of ACL-2016 System Demonstrations*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 43–48. [Online]. Available: http://anthology.aclweb.org/P16-4008

[16] C. Servan, N.-T. Le, N. Q. Luong, B. Lecouteux, and L. Besacier, "An Open Source Toolkit for Word-level Confidence Estimation in Machine Translation," in *The 12th International Workshop on Spoken Language Translation (IWSLT'15)*, Da Nang, Vietnam, Dec. 2015. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01244477

[17] L. Besacier, B. Lecouteux, N. Q. Luong, K. Hour, and M. Hadjsalah, "Word confidence estimation for speech translation," in *Proceedings of The International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, USA, December 2014.

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

[19] 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.

[20] A. Laurent, N. Camelin, and C. Raymond, "Boosting bonsai trees for efficient features combination : application to speaker role identification," in *Interspeech*, 2014.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "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.

[26] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 evaluation campaign," in *In proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT)*, December 2012.

[27] 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.