Word Confidence Estimation For Speech Translation
Laurent Besacier, Benjamin Lecouteux, Ngoc Quang Q Luong, K Hour, Marwa Hadj Salah

To cite this version:
Laurent Besacier, Benjamin Lecouteux, Ngoc Quang Q Luong, K Hour, Marwa Hadj Salah. Word Confidence Estimation For Speech Translation. IWSLT, 2014, Lake Tahoe, United States. hal-02088818

HAL Id: hal-02088818
https://hal.science/hal-02088818
Submitted on 3 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Introduction

Word Confidence Estimation for machine translation or automatic speech recognition 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 task involving both ASR and MT. This research work is possible because we built a specific corpus which is first presented.

A database for WCE evaluation in spoken language translation

Starting point : an existing MT Post-edition corpus

For a Fr-En translation task, we used our SMT system to obtain the translation hypothesis for 10,881 source sentences taken from news corpuses of the WMT evaluation campaign (2006-2010).
Post-editions were obtained from non professional translators using a crowdsourcing platform.
Word label setting for WCE was done using TERp-A toolkit.

Reducing the categories 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.

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.

TABLE: Example of WCE label setting using TERp-A

Augmenting the corpus with speech recordings and transcripts

We record the utterances of PE corpus test to augment the corpus with speech inputs.
Each of the 881 sentences was uttered by 3 speakers, leading to 2643 speech recordings (5h) : 15 speakers (9 women and 6 men).
ASR system based on KALDI toolkit with a 3-gram LM trained on the French ESTER corpus and French Gigaword (vocabulary size is 55k). SGMM acoustic models are trained on the ESTER corpus.
Post-processing was needed at the output of the ASR system in order to match requirements of standard input for machine translation.

The output of the ASR system, scored against the src-ref reference is 26.6% WER (these news contain a lot of foreign named entities).

TABLE: Overview of our post-edition corpus for SLT

Corpus available for download on github.com/besacier/WCE-SLT-LIG.

Obtaining labels in order to evaluate WCE for SLT:

The ASR output (src-asr) was translated by the SMT system (tgt-slt, a degraded version of tgt-mt).
We re-used the post-editions obtained from the text translation task (tgt-pe), to infer the quality (GB) labels of our speech translation output tgt-sl.
The word label setting for WCE is done using TERp-A toolkit between tgt-sl and tgt-pe.

TABLE: Example of quintuplet with associated labels

Word confidence for a speech translation task

WCE for speech transcription

7 features (F-Word : F-3g, F-back, F-alt, F-post; F-dur; F-post) :
Acoustic features : acoustic distortions between the hypothesis and the best phonetic sequence (F-dur).
Graph features : extracted from the word confusion networks (number of alternative (F-Alt) paths in the word section, and the posterior probability (F-post)).
Lexical features : probabilities provided by the language model (3-gram LM). We use the word itself (F-word), the 3-gram probability (F-3g) and the back-off behavior (F-back).
Lexical Features: words' Part-Of-Speech (POS), sequence of POS of all its aligned source words ; POS bigram (gram) backward sequences ; punctuation ; proper name ; numerical.
Syntactic Features : null link ; constituent label ; depth in the constituent tree.
Semantic Features : number of word senses in WordNet.
Graph topology : number of alternative paths in the confusion set, maximum and minimum values of posteriors.
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 w3, if the sequence w1,w2,w3 appears in the target LM but the sequence w2,w3 does not, the n-gram value for w3 will be 3.
Lexical Features : words' Part-Of-Speech (POS) ; sequence of POS of all its aligned source words ; POS bigram (gram) backward sequences ; punctuation ; proper name ; numerical.
Syntactic Features : null link ; constituent label ; depth in the constituent tree.
Semantic Features : number of word senses in WordNet.

WCE for machine translation

We employ CRFs as our machine learning method, with WAPITI toolkit, to train the WCE model. 25 major feature types : Target Side : target word ; bigram (gram) backward sequences ; number of occurrences.
Source Side : source word(s) aligned to the target word.
Alignment Context: the combinations of the target (source) word and all aligned source (target) words in the window : i-2.
Word posterior probability.

Pseudo-reference (Google Translate) : Does the word appear in the pseudo reference or not?

Joint estimation of word confidence for a speech translation task

<table>
<thead>
<tr>
<th>Feature</th>
<th>ASR feat</th>
<th>MT feat</th>
<th>0.5MT+0.5ASR feat</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-dur</td>
<td>0.0000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>F-3g</td>
<td>0.0000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>F-back</td>
<td>0.0000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>F-alt</td>
<td>0.0000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>F-post</td>
<td>0.0000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
</tbody>
</table>

TABLE: 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.

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

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