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

Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France http://www.liglab.fr/
GETALP Team http://getalp.imag.fr

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.

This corpus 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 made available. The rest of the paper illustrates how such a corpus 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 or it could be useful in interactive speech-to-speech translation scenarios.

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 corpora 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.
- 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.
- 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.

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: Example of WCE label setting using TERp-A

Augmenting the corpus with speech recordings and transcripts

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

Final corpus statistics and web link for download

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: Overview of our post-edition corpus for SLT

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

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: Exemple of quintuplet with associated labels

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 (G,B) labels of our speech translation output *tgt-slt*. The word label setting for WCE is done using TERp-A toolkit between *tgt-slt* and *tgt-pe*.

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: 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.

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)).
- Linguistic 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 : word's Part-Of-Speech (F-POS) are computed using tree-tagger for French.

We use bonzaiboost algorithm, the classifier is trained on BREF 120 corpus (about 1M word examples). Each word is tagged as correct or not correct, according to the reference.

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 (trigram) 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 ± 2.
- Word posterior probability.
- Pseudo-reference (Google Translate) : Does the word appear in the pseudo reference or not ?
- 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 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 ; constituent label ; depth in the constituent tree.
- Semantic Features : number of word senses in WordNet.

Joint estimation of word confidence for a speech translation task

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: 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.

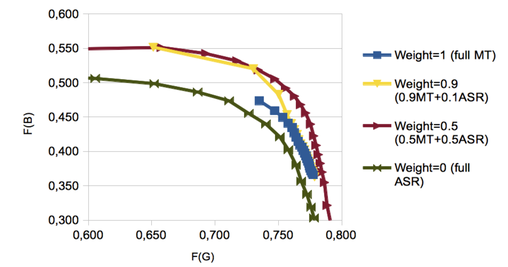

FIGURE: WCE performance (F(B) vs F(G) of different WCE methods - for SLT - for different decision thresholds varying from 0.5 to 0.9).

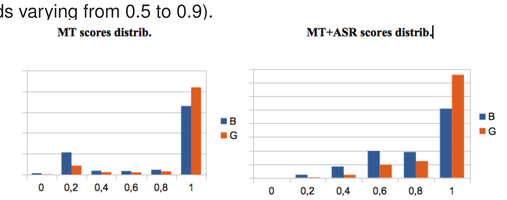

FIGURE: Evolution of the WCE scores distribution from MT features to MT+ASR features