Better Evaluation of ASR in Speech Translation Context Using Word Embeddings

Ngoc-Tien Le, Christophe Servan, Benjamin Lecouteux and Laurent Besacier

LIG - Univ. Grenoble Alpes

firstname.lastname@imag.fr

Abstract

This paper investigates the evaluation of ASR in spoken language translation context. More precisely, we propose a simple extension of WER metric in order to penalize differently substitution errors according to their context using word embeddings. For instance, the proposed metric should catch near matches (mainly morphological variants) and penalize less this kind of error which has a more limited impact on translation performance. Our experiments show that the correlation of the new proposed metric with SLT performance is better than the one of WER. Oracle experiments are also conducted and show the ability of our metric to find better hypotheses (to be translated) in the ASR N-best. Finally, a preliminary experiment where ASR tuning is based on our new metric shows encouraging results. For reproductible experiments, the code allowing to call our modified WER and the corpora used are made available to the research community.

Index Terms: Spoken Language Translation, Automatic Speech Recognition, Evaluation Metrics, Correlations, Word Embeddings

1. Introduction

In spoken language translation (SLT), the ability of Word Error Rate (WER) metric to evaluate the real impact of the ASR module on the whole SLT pipeline is often questionned. This was investigated in past studies where researchers tried to propose a better evaluation of ASR in speech translation scenarios. [1] investigated how SLT performed as they changed speech decoder parameters. It was shown that sub-optimal WER values could give comparable BLEU scores at faster decoding speeds. The authors of [2] analyzed ASR error segments that have a high negative impact on SLT performance and demonstrated that removing such segments prior to translation can improve SLT. The same year, [3] proposed a Phonetically-Oriented Word Error Rate (POWER) for speech recognition evaluation which incorporates the alignment of phonemes to better trace the impact of Levenshtein error types in speech recognition on downstream tasks (such as information retrieval, spoken language understanding, speech translation, etc.). Moreover, the need to evaluate ASR speech recognition when its output is used by human subjects (predict how useful that ASR output would be to humans) was also highlighted by [4]. Finally, some authors [5] proposed an end-to-end BLEU-oriented global optimization of ASR system parameters in order to improve translation quality. However, such an end-to-end optimization is not always possible in practical applications where a same ASR system is designed for several downstream uses. Thus, we believe that a better evaluation of the ASR module itself should be investigated.

Contribution This paper rests upon the above papers as well as on the former research of [6] who noticed that many ASR substitution errors (the most frequent type of ASR error) are due to slight morphological changes (such as plural/singular substitution), limiting the impact on SLT performance. Thus, the current WER metric – which gives the same weight to any subsitution – is probably sub-optimal for evaluating ASR module in a SLT framework. We propose a simple extension of WER in order to penalize differently substitution errors according to their context using word embeddings. For instance, the proposed metric should penalize less morphological changes that have a smaller impact on SLT. We specifically extend our existing French-English corpus for SLT evaluation and shows that the new proposed metric is better correlated with SLT performance. Oracle experiments are also conducted to show the ability of our metric to find better hypotheses (to be translated) in the ASR N-best. Finally, we propose a preliminary experiment where ASR tuning is based on our new metric. For reproductible experiments, code allowing to call our modified WER and corpora used are made available to the research community.

Outline The rest of the paper goes simply as follows: section 2 summarizes related works on evaluation metrics that use word embeddings. Section 3 presents our modified WER metric which allows to consider near matches in substitution errors. Section 4 details the experimental settings and section 5 presents our results. Section 6 concludes this work.

2. Related works on evaluation metrics using word embeddings

Word embeddings are a representation of words in a continuous space. Mikolov and al. [7] have shown that these vector representations could be useful to detect near matches (like syntactic variants or synonyms). For this work, we decided to choose the representation proposed by [8] and implemented in the toolkit MultiVec [9]. The use of word embeddings has grown since the work done by Mikolov [8], especially in Natural Language Processing (NLP). Tasks such as machine translation [10], information retrieval [11] and many others, use continuous word representations. As far as we know, only few works used word embeddings for evaluation in NLP. One of them is the paper recently published by [12] which extends ROUGE, a metric used in text summarization. Concerning Machine Translation, [13] proposed a metric (for WMT 2015 *metrics* shared task) that represents both reference and translation hypotheses using a dependency Tree-LSTM and predicts the similarity score based on a neural network. In the same workshop, [14] used document embeddings for predicting MT adequacy. These two latter works are close to what we propose. However, they both rely on the training of the metric itself which questions its portability to evaluation on other domains / tasks. In our work, we propose to use word embeddings that are trained once and for all on a general corpus.

3. *WER* with embeddings (*WER–E*)

The Word Error Rate is the main metric applied to Automatic Speech Recognition evaluation. Its estimation is based on the Levenshtein distance, which is defined as the minimum number of editing steps needed to match an hypothesis and a reference.

	un	nord	westphalie	un	d'	engagement	parmi	de	nation	souveraine
un	0	1	2	3	4	5	6	7	8	9
ordre	1	1	2	3	4	5	6	7	8	9
westphalien	2	2	2	3	4	5	6	7	8	9
d'	3	3	3	3	3	4	5	6	7	8
engagements	4	4	4	4	4	4	5	6	7	8
parmi	5	5	5	5	5	5	4	5	6	7
des	6	6	6	6	6	6	5	5	6	7
nations	7	7	7	7	7	7	6	6	6	7
souveraines	8	8	8	8	8	8	7	7	7	7
Alignment:	A	I	S	S	A	S	A	S	S	S
Cost:	0	1	1	1	0	1	0	1	1	1

Table 1: Example (in French) of the Word Error Rate estimation between a hypothesis (on the top) and a reference (on the left).

3.1. Running exemple

In table 1, we compare an hypothesis (on the top) and a reference (on the left): the score is defined as the lowest-cost alignment path (in grey) from the beginning of both sentences (top left corner) to the end of both sentences (on the lower-right corner). The intensity of the colour in the alignment path indicates the match level: lighter grey for matches, mid-dark grey for *substitutions* and dark grey for *insertions* and *deletions*. The score sums the number of *insertions*, *deletions* and *substitutions*. Then, this sum is normalized by the length of the reference. In our example, the *WER* is 78% (0.78).

3.2. Adding word embeddings

The main drawback of *WER* is that it does not gives credit to near matches. For instance, in table 1, the hypothesis contains the word "souveraine", which is close to the word "souveraines" in the reference. Both are morphological variants of a same word and WER considers this difference as a *Substitution*, while their cosine distance in the continuous space is only 0.43.

	un	nord	westphalie	un	d'	engagement	parmi	de	nation	souveraine
un	0	1	2	3	4	5	6	7	8	9
ordre	1	1.01	2.07	2.93	4.15	4.89	6.07	7.03	8.05	9.01
westphalien	2	1.79	1.73	2.83	3.93	5.38	5.80	6.90	7.75	8.85
d'	3	3.05	2.97	2.21	2.83	3.83	4.83	5.83	6.83	7.83
engagements	4	3.94	4.02	4.15	3.41	3.30	5.01	5.91	6.92	7.81
parmi	5	4.77	4.80	5.13	5.15	4.61	3.30	4.30	5.30	6.30
des	6	6.04	5.85	5.80	5.61	6.24	4.30	3.64	5.49	6.12
nations	7	6.87	6.83	6.77	6.85	6.55	5.30	5.26	4.42	6.43
souveraines	8	7.92	7.71	7.99	7.71	7.82	6.30	6.15	6.10	**4.85**
Alignment:	A	I	S	S	A	S	A	S	S	S
Cost:	0	1	1.07	0.75	0	0.47	0	0.35	0.78	0.43

Table 2: WER-E estimation with word embeddings. *Substitution* score is replaced by a cosine distance, without questionning the best alignment.

Our main idea is to find a way to include near matches in the metric without using lexico-semantic data such as Wordnet. Since word embeddings can model syntactic and semantic proximity [8, 15], we use them to estimate a cosine similarity between two words in a *substitution*. This cosine similarity (S_c in [-1,1]) is used to compute a cosine distance (D_c) (see equation 1). The *substitution* score (0 or 1) is replaced by the cosine distance between two words (continuous value in [0,2]).

$$D_c(W_1, W_2) = 1 - S_c(W_1, W_2) \tag{1}$$

From this, two variants of the metric are possible. Firstly, in table 2, we apply the *WER* alignment algorithm with classical *substitution* cost (we do not modify the alignment path of table 1) and we replace only the *substitution* scores by the cosine distance. We call it "WER *with embeddings*" (*WER-E*). Secondly, in table 3, we propose to replace *substitution* cost by the cosine distance to compute the best alignment path. We call this last WER variant "WER *soft*" (*WER-S*).

In the first case (table 2), we can observe a *WER-E* score (54%) lower than the classical *WER* estimation (78%). Since we do not question the alignment path in this case, we do not obtain the lowest score possible. The second case, presented in table 3, enables us to get another alignment path, and thus gets the lowest score possible (53%).

This new feature takes into account near matches between words. For instance, words "westphalie" and "westphalien" are close enough to have a low distance. In the alignment proposed in table 3, the alignment changed and we got a lower score.

	un	nord	westphalie	un	d'	engagement	parmi	de	nation	souveraine
un	0	1	2	3	4	5	6	7	8	9
ordre	1	1.01	2.01	2.93	3.93	4.89	5.89	6.89	7.89	8.89
westphalien	2	1.79	1.74	2.74	3.74	4.74	5.74	6.72	7.61	8.61
d'	3	2.79	2.74	2.21	2.74	3.74	4.74	5.74	6.74	7.74
engagements	4	3.79	3.74	3.21	3.42	3.21	4.21	5.21	6.21	7.21
parmi	5	4.77	4.65	4.21	4.21	4.21	3.21	4.21	5.21	6.21
des	6	5.77	5.65	5.21	4.68	5.21	4.21	3.55	4.55	5.55
nations	7	6.77	6.57	6.21	5.68	5.63	5.21	4.55	4.34	5.34
souveraines	8	7.77	7.57	7.21	6.68	6.63	6.21	5.55	5.34	4.76
Alignment:	A	S	S	I	A	S	A	S	S	S
Cost:	0	1.01	0.73	1	0	0.47	0	0.35	0.78	0.43

Table 3: WER-S estimation with word embeddings. *Substitution* score is replaced by a cosine distance **and** we recalculate the best alignment.

4. Dataset and ASR, MT, SLT systems

For the experiments of this paper, we have extended our corpus presented in [16]. This corpus, available on a *github* repository[1] contained initially 2643 French speech utterances (news domain) x_f for which 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. We recently added 4050 new sentences of the same (news) domain in our corpus (our *github* repository has been updated with this new data). The initially available corpus (2643 utterances) will be refered to as *dev* set in the rest of the paper while the recently recorded part (4050 utterances) will be refered to as *test* set in the rest of the paper. For ASR output, the N-best lists (N=1000) were also generated for each utterance.

4.1. ASR system

To obtain the speech transcripts (f_{hyp}), we built a French ASR system based on KALDI toolkit [17]. It is trained using several corpora (ESTER, REPERE, ETAPE and BREF 120) representing more than 600 hours of transcribed speech. CD-DNN-HMM acoustic models are trained (43 182 context-dependent

[1] https://github.com/besacier/WCE-SLT-LIG/

REF ASR	ce serait intéressant de voir un ordinateur présentant ce même système	WER	WER-E	WER-S
OptWER	ce **sera** intéressant de voir un ordinateur présentant ce même système	9.09	2.43	2.43
OptWER-E	ce **serait** intéressant de voir un ordinateur présentant ce même système	0.00	0.00	0.00
REF SLT	it would be interesting to see a computer with this same system	TER	SentBLEU	METEOR
OptWER - SLT	this **will** be interesting to see a computer with the same system	33.33	62.63	49.33
OptWER-E - SLT	it **would** be interesting to see a computer with the same system	16.67	79.11	92.73

REF ASR	en bref ils craignent que tous les sacrifices entrepris pour stabiliser les prix aient été vains	WER	WER-E	WER-S
OptWER	en bref **il craignait** que tous les sacrifices ces entreprises pour stabiliser les prix et était vingt	43.75	34.65	33.26
OptWER-E	en bref **ils craignent** que tous les sacrifices ces entreprises pour stabiliser les prix et était vingt	31.25	26.80	25.41
REF SLT	in short they fear that all the sacrifices made to stabilize prices have been fruitless	TER	SentBLEU	METEOR
OptWER - SLT	in short **it feared** that all the sacrifices these companies to stabilise prices and was 20	60.00	26.22	34.84
OptWER-E - SLT	in short **they fear** that all the sacrifices these companies to stabilise prices and was 20	46.67	50.44	40.08

Table 4: ASR and SLT examples (explanations given in *section 5.5*)

states) and the 3-gram language model is learned on French ESTER corpus [18] as well as on French Gigaword (vocabulary size is 55k). The ASR system's LM weight parameter is tuned through WER on the *dev* corpus. The output of our ASR system, scored against the f_{ref} reference is 21.92% WER on *dev* set and 17.46% WER on *test* set. This WER may appear as rather high according to the task (transcribing read news). A deeper analysis shows that these news contained a lot of foreign named entities, especially in our *dev* set. This part of the data is extracted 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.

4.2. SMT system

We used *moses* phrase-based translation toolkit [19] to translate French ASR into English (e_{hyp}). This medium-sized system was trained using a subset of data provided for IWSLT 2012 evaluation [20]: 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 [21]). Table 4 gives 2 examples of SLT output obtained. Table 5 summarizes baseline ASR, MT and SLT performances obtained on our corpora. We score translations obtained with the following automatic metrics: TER [22], BLEU [23] and METEOR [24] using post-edition references (e_{ref}).

Tasks	*metrics*	ASR Ref.	ASR 1-best
dev	*WER*	–	21.92
	TER	38.84	55.64
	BLEU	43.05	30.81
	METEOR	40.73	34.02
test	*WER*	–	17.46
	TER	45.64	58.70
	BLEU	44.71	34.27
	METEOR	39.10	34.27

Table 5: Baseline ASR, MT and SLT performance on our *dev* and *test* sets - translations are scored w/o punctuation

5. Experiments and results

This section first presents the results obtained in ASR, according to our new metrics. Then, we analyze the correlation of the ASR metrics (WER, WER-E, WER-S) with SLT performances. After that, Oracle experiments are conducted to compare the ASR metrics in their ability to find (before translation) promising hypotheses in the ASR N-best. Finally, a preliminary experiment where ASR tuning is based on our new metric is proposed. For all the experiments, the MT system never changes and is the one described in section 4.

Tasks	*metrics*	ASR 1-best	Oracle from N-best WER	WER-E	WER-S
dev	*WER*	21.92	12.01	12.16	12.15
	WER-E	18.10	10.45	9.98	10.04
	WER-S	17.41	10.19	9.79	9.75
test	*WER*	17.46	7.38	7.53	7.52
	WER-E	13.13	5.86	5.43	5.48
	WER-S	12.53	5.65	5.29	5.25

Table 6: Speech Recognition (ASR) performance - ASR Oracle is obtained from 1000-best list by selecting hypothesis that minimizes WER, WER-E or WER-S

5.1. ASR results

Table 6 presents the performances obtained by the ASR system described in section 4. The columns correspond to four settings: the best output according to the ASR system, and three oracles extracted from the N-best list. The oracle ASR performances are obtained by sorting the N-best hypotheses according to WER, WER-E or WER-S. The results show that the oracle hypotheses selected by WER, WER-E and WER-S can be different. In other words, optimizing the ASR according to the new metrics proposed can degrade WER but improve WER-E or WER-S. In this case, better ASR outputs in term of near matches are selected. Overall, whatever the metric used, Oracle hypotheses contain approximately 50% of the initial errors found in the 1-best.

5.2. Correlation between ASR metrics and SLT performance

In this section, we investigate if our new metrics WER-E and WER-S are better correlated with speech translation (SLT) performance. Table 7 shows the correlation (*Pearson*) between ASR metrics (WER, WER-E or WER-S) and SLT performances (TER, BLEU, METEOR). Since BLEU and METEOR are not very efficient to evaluate translations at the sentence level, we

Tasks	*metrics*	Pearson Correlation WER	WER-E	WER-S
dev	*TER*	0.732	0.767	**0.773**
	BLEU	-0.677	-0.708	**-0.710**
	METEOR	-0.753	**-0.799**	-0.797
tst	*TER*	**0.457**	**0.457**	0.441
	BLEU	-0.624	**-0.661**	-0.606
	METEOR	-0.672	**-0.692**	-0.678

Table 7: Pearson Correlation between ASR metrics (WER, WER-E or WER-S) and SLT performances (TER, BLEU, METEOR) - each point measured on blocks of 100 sentences

decided to group our sentences by blocks of 100 (in order to have relevant measure points for correlation analysis). We end up with 27 blocks on *dev* and 41 blocks on *test* for evaluating correlation. The reading of the TER score is "the lower the better", and BLEU and METEOR are "the higher the better" which explains the different signs of the correlation values. The results show clearly a better correlation of the proposed metrics (WER-E and WER-S) with SLT performances, compared to classical WER. Also, we notice that all ASR metrics are better correlated with METEOR (itself known to be better correlated with human judgements), while ASR metrics are less correlated with BLEU.

We finally investigate the Spearman correlation coefficient for which we observe the same trend (results not reported here).

5.3. Oracle analysis

In this section, we verify if the hypotheses selected by WER-E and WER-S are more promising for translation. Our Oracle analysis is presented in Table 8. Similarly to Table 6, the columns correspond to four settings: the best output according to the ASR system is translated, and three oracles are scored by translating the most promising hypotheses according to WER, WER-E or WER-S. Even if there are not big differences in SLT performance, the results show the ability of our metric to find slightly better hypotheses (to be translated) in the ASR N-best. For instance, when the WER-S score is used to select the best ASR hypothesis, the TER, BLEU and METEOR are improved by respectively 0.18, 0.12, and 0.06 points on the *dev* corpus.

Tasks	*metrics*	ASR 1-best	Oracle from N-best WER	WER-E	WER-S
dev	*TER*	55.64	50.62	50.52	**50.45**
	BLEU	30.81	35.29	35.37	**35.41**
	METEOR	34.02	36.37	36.42	**36.44**
test	*TER*	58.70	54.13	**54.01**	54.03
	BLEU	34.27	39.34	**39.43**	39.42
	METEOR	34.27	36.55	**36.64**	**36.64**

Table 8: Speech Translation (SLT) performances - Oracle is obtained from 1000-best list by translating hypothesis that minimizes WER, WER-E or WER-S

We also analyzed how often the *Oracle (according to WER-E)* system obtain better results at the sentence level compared to the *Oracle (according to WER)*. Table 9 shows this comparison for the three MT metrics (TER, sentenceBLEU and METEOR). Even if we logically observe a majority of ties where *Oracle (according to WER-E)* and *Oracle (according to WER)* lead to the same SLT output, for the other cases the analysis shows a preference of the translation metrics for the *Oracle (according to WER-E)*. This result confirms the trend observed in table 8.

Tasks	Comparison	TER	BLEU	METEOR
Dev	O. WER-E best	**255**	**310**	**321**
	O. WER best	190	271	315
	Ties	2198	2062	2007
Test	O. WER-E best	**341**	**451**	**510**
	O. WER best	264	381	399
	Ties	3445	3218	3141

Table 9: Comparison of SLT performances of the *Oracle WER* vs. the *Oracle WER-E* by counting the number of sentences which obtain a better MT score according to TER, Sentence BLEU and METEOR.

Tasks	*metrics*	ASR optimized with WER	ASR optimized with WER-E
dev	*TER*	55.64	**55.52**
	BLEU	30.81	**30.84**
	METEOR	**34.02**	34.00
test	*TER*	58.71	**58.56**
	BLEU	34.27	**34.38**
	METEOR	**34.27**	34.26

Table 10: Speech Translation (SLT) scores obtained with 2 ASR systems optimized with WER or WER-E

5.4. ASR optimization for SLT

This section investigates if the tuning of an ASR system using the new metrics proposed can lead to real (and not oracle) improvements. This experiment is preliminary since we only optimize the LM weight parameter (to minimize WER or WER-E[2]) on the *dev* corpus.

The results are given in table 10 but they are not very convincing: we observe small gains for TER and BLEU evaluation but not improvement of METEOR. Our explanation is that there were too few free parameters investigated to tune the ASR system. In addition, translation evaluation metrics are themselves unperfect to evaluate translation quality. The next section proposes to analyze a few translation examples to better understand the differences of both SLT systems.

5.5. Translation exemples

In table 4 are presented some translation examples related to the ASR optimization. We can observe in these examples that both ASR systems (*OptWER* and *OptWER-E*) are very close. For instance, in the first example, the ASR hypothesis is different only on one word ("sera" vs. "serait"). Both are the same verb at the right agreement with the pronoun but not at the same tense. These are two examples where the ASR optimized according to WER-E lead to better translation (SLT) hypotheses than WER. What it means is simply the fact that ASR system is optimized according to a metric which penalizes less substitutions between "morphologically similar" words. We believe that for optimizing ASR systems along a larger number of meta-parameters, the modififed metrics proposed in this paper could be even more useful.

6. Conclusions

We proposed an extension of WER in order to penalize differently substitution errors according to their context using word embeddings. Our experiments, made on a French-English speech translation task, have shown that the new proposed metric is better correlated with SLT performance. Oracle experiments have also shown a trend: the ability of our metric to find better hypotheses (to be translated) in the ASR N-best. This opens possibilities to optimize ASR using metrics clever than WER. For reproductible experiments, code allowing to call our modified WER is made available on *github*[3].

7. Acknowlegements

This work was partially founded by the French National Research Agency (ANR) through the KEHATH Project.

8. References

[1] P. R. Dixon, A. Finch, C. Hori, and H. Kashioka, "Investigation on the effects of ASR tuning on speech trans-

[2] WER-S lead to the same optimized ASR system than WER-E

[3] https://github.com/cservan/tercpp-embeddings

lation performance," in *The proceedings of the International Workshop on Spoken Language Translation (IWSLT 2011)*, San Francisco, CA, USA, December 2011.

[2] F. Bechet, B. Favre, and M. Rouvier, """speech is silver, but silence is golden": improving speech-to-speech translation performance by slashing users input," in *Proceedings of Interspeech 2015*, Dresden, Germany, September 2015.

[3] N. Ruiz and M. Federico, "Phonetically-oriented word error alignment for speech recognition error analysis in speech translation," in *IEEE 2015 Workshop on Automatic Speech Recognition and Understanding*, December 2015.

[4] B. Favre, K. Cheung, S. Kazemian, A. Lee, Y. Liu, C. Munteanu, A. Nenkova, D. Ochei, G. Penn, S. Tratz, C. Voss, and F. Zeller, "Automatic Human Utility Evaluation of ASR Systems: Does WER Really Predict Performance?" in *Proceedings of Interspeech 2013*, Lyon, France, August 2013.

[5] X. He, L. Deng, and A. Acero, "Why word error rate is not a good metric for speech recognizer training for the speech translation task?" in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011.

[6] D. Vilar, J. Xu, L. F. D'Haro, and H. Ney, "Error analysis of statistical machine translation output," in *Proceedings of LREC 2006*, Genoa, Italy, May 2006.

[7] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, GA, USA, June 2013.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *The Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013*, Scottsdale, AR, USA, May 2013.

[9] A. Bérard, C. Servan, O. Pietquin, and L. Besacier, "MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP," in *The 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, May 2016.

[10] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 2014.

[11] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "A latent semantic model with convolutional-pooling structure for information retrieval," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, November 2014.

[12] J. Ng and V. Abrecht, "Better summarization evaluation with word embeddings for ROUGE," in *In The Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, September 2015.

[13] R. Gupta, C. Orasan, and J. V. Genabith, "Machine translation evaluation using recurrent neural networks," in *Proceedings Workshop on Machine Translation (WMT), Metrics Shared Task*, Lisbon, Portugal, September 2015.

[14] M. Vela and L. Tan, "Predicting machine translation adequacy with document embeddings," in *Proceedings Workshop on Machine Translation (WMT), Metrics Shared Task*, Lisbonne, Portugal, 2015.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds. Curran Associates, Inc., 2013.

[16] 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, CA, USA, December 2014.

[17] 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*, December 2011.

[18] 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.

[19] 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, Juin 2007.

[20] 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.

[21] 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, July 2010.

[22] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of association for machine translation in the Americas*, 2006.

[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[24] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics*, 2005.