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

In this paper we try to correlate the scores of ASR outputs and its translation scores.

- Max 4+1 pages.

Index Terms: ASR, SMT, Automatic Metrics, 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.

[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 speech translation.

While some authors like [4] proposed an end-to-end BLEU-oriented global optimization of ASR system parameters in order to improve translation quality, such an end-to-end optimization is not always possible in some practical applications and we believe that the need for a better evaluation of the single ASR block remains. Moreover, a same ASR system may be designed for several downstream uses (such as information retrieval, spoken language understanding, speech translation, etc.) which prevents the use of joint end-to-end optimization.

Finally, [5] also highlighted 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).

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 more limited impact on SLT. We specifically extend our existing French-English corpus for SLT evaluation and show that the correlation of the new proposed metric with SLT performances is better than the one of WER. Oracle experiments are also made to show that our modified WER is better to find the best hypothese (to be translated) from an ASR N-best list.

For reproductible experiments, the code allowing to call our modified WER and the corpora used are made available to the research community.

Outline

Laurent will do it.

2. Related works on evaluation metrics using word embeddings

2.1. Word embeddings

Word embeddings are a representation of words in a continuous space. Mikolov and al.[7] have shown that these vectors could be useful to detect near matches (like syntactic variants or synonyms). We chose to use the representation proposed by [8] and its toolkit word2vec[1]. This toolkit uses two different models: the Continuous Bag-of-Words (*CBOW*) and the Continuous Skip-gram model (*Skip-gram*). Theses two models that are based on a feedforward Neural Network Language Model (*NNLM*) [9].

In the *CBOW* model, the hidden layer of the *NNLM* is removed and the projection layer is shared for all words. To predict a specified word, the model uses the words around within a determined window size. All the words are projected into the same position and the information related to original positions of words in the history is lost. This is why this architecture is called a Bag-of-Word (*BoW*) model but, unlike a standard *BoW*, it uses a continuous representation of the context.

The *Skip-gram* model is the reverse of the *CBOW* model: instead of predicting the current word based on the surrounding words, the model uses the current word as an input to predict words within a defined window (before and after the current word).

Word embeddings can be learned by using one of these two models associated to some other parameters like the size of the vocabulary, the window size around the word considered, the softmax activation function associated to the network, *etc.* The result of the learning phase is a set of vectors in a continuous space, which corresponds to a vocabulary.

2.2. Using word embeddings in evaluation metrics

The use of word embeddings have grown since the work done by Mikolov [8, 10], especially in Natural Language Processing

[1] https://code.google.com/p/word2vec/

(NLP). Tasks such as machine translation [11], information retrieval [12], question answering systems [13] and many others, use continuous word representations.

As far as we know, only few works used word embeddings for evaluation. One of them is the paper recently published by [14] which extends ROUGE, a metric used in text summarization. Concerning Machine Translation, [15] proposed a metric (for WMT 2015 *metrics* shared task) that represents both reference and translation hypothesis using a dependency Tree-LSTM and predicts the similarity score based on a neural network. In the same workshop, [16] used document embeddings for predicting MT adequacy. These two latter works are close to what we propose however, the 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 general corpora.

In addition, another work, close to our approach, was proposed by [17], who learn bilingual word embeddings to detect semantic word similarities for word alignement and add this information as an additional feature to a phrase-based machine translation system (whereas we propose to use word embeddings in an automatic metric). Finally, [18] uses a continuous representation of the semantic proximity between two sentences in a metric: the Latent Semantic Indexing [19] which is an alternative to the use of word embeddings to detect near matches between words.

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.

		the	scientist	said	that	much	more	thorough	searches	were	necessary
	0	1	2	3	4	5	6	7	8	9	10
the	1	0	1	2	3	4	5	6	7	8	9
scientist	2	1	0	1	2	3	4	5	6	7	8
said	3	2	1	0	1	2	3	4	5	6	7
far	4	3	2	1	1	2	3	4	5	6	7
more	5	4	3	2	2	2	2	3	4	5	6
research	6	5	4	3	3	3	3	3	4	5	6
was	7	6	5	4	4	4	4	4	4	5	6
needed	8	7	6	5	5	5	5	5	5	5	**6**
Alignment:		A	A	A	I	S	A	I	S	S	S
Cost:		0	0	0	1	1	0	1	1	1	1

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

3.0.1. 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.

$$WER = \frac{Insertions + Deletions + Substitutions}{length\ of\ the\ reference} \quad (1)$$

The score sums the number of words added (*Insertions*) in the hypothesis, the words missing (*Deletions*) in it, and the wrong words aligned with the reference (*Substitutions*). Then, this sum is normalized by the length of the reference (see equation 1).

In our example, the *WER* is equal to 0.75 (six errors divided by a reference length of eight).

3.0.2. 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 "needed", which is close to the word "necessary" in the reference. Both have the same meaning but since the surface forms of the words are not the same, the metric considers this difference as a full error (a *Substitution* error - while their cosine distance in the continuous space is only 0.22).

		the	scientist	said	that	much	more	thorough	searches	were	necessary
	0	1	2	3	4	5	6	7	8	9	10
the	1	0	1	2	3	4	5	6	7	8	9
scientist	2	1	0	1	2	3	4	5	6	7	8
said	3	2	1	0	1	2	3	4	5	6	7
far	4	3	2	1	0.68	1.33	2.73	4.01	5.11	5.71	7.08
more	5	4	3	2	1.77	1.17	1.33	2.33	3.33	4.33	5.33
research	6	5	4	3	3.06	2.69	2.04	2.10	3.20	4.53	5.21
was	7	6	5	4	3.68	4.02	3.77	2.99	3.11	3.62	5.88
needed	8	7	6	5	4.84	4.60	4.65	4.58	3.98	4.33	**3.84**
Alignment:		A	A	A	I	S	A	I	S	S	S
Cost:		0	0	0	1	0.33	0	1	0.87	0.42	0.22

Table 2: The *WER* estimation with word embeddings: the *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 collecting more references nor using linguistic data. Word embeddings can model syntactic and semantic proximity [8, 10]. It enables us to estimate a cosine similarity between two words, considering their representation in a continuous space. This cosine similarity (S_c) will be used to get its opposite: the cosine distance (D_c), which is more meaningful for the *WER* estimation (see equation 2). Consequently, we replace the *Substitution* score by the cosine distance between two words.

$$D_c(W_1, W_2) = 1 - S_c(W_1, W_2) \quad (2)$$

We propose to observe this replacement in two ways. Firstly, in table 2, we apply the *WER* algorithm with the classical *Substitution* cost (i.e.: we do not modify the alignment path presented in table 1) and we replace only the *Substitution* scores by Sub_{WE} like presented in the equation 3. Secondly, in table 3, we propose to replace the *Substitution* error in the Levenshtein algorithm by Sub_{WE} to compute the best alignment path.

$$Sub_{WE}(W_1, W_2) = \begin{cases} 1,\ if\ W_1\ or\ W_2\ are\ OOV \\ D_c(W_1, W_2),\ otherwise \end{cases} \quad (3)$$

In the first case (table 2), we can observe a *WER* score lower than the classical *WER* estimation. 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 (differences are in bold).

This new feature seems to capture the near match between words. For instance, the words "needed" and "necessary" have

		the	scientist	said	that	much	more	thorough	searches	were	necessary
	0	1	2	3	4	5	6	7	8	9	10
the	1	0	1	2	3	4	5	6	7	8	9
scientist	2	1	0	1	2	3	4	5	6	7	8
said	3	2	1	0	1	2	3	4	5	6	7
far	4	3	2	1	0.68	1.33	2.33	3.33	4.33	5.33	6.33
more	5	4	3	2	1.68	1.17	1.33	2.33	3.33	4.33	5.33
research	6	5	4	3	2.68	2.17	2.04	2.10	3.10	4.10	5.10
was	7	6	5	4	3.68	3.17	3.04	2.99	3.11	3.53	4.53
needed	8	7	6	5	4.68	4.17	3.80	3.85	3.98	4.33	**3.74**
Alignment:		A	I	A	I	S	A	**S**	**I**	**S**	**S**
Cost:		0	1	0	1	0.425	0	**0.77**	**1**	**0.42**	**0.22**

Table 3: The *WER* estimation with word embeddings: the *Substitution* score is replaced by a cosine distance **and** we recalculate the best alignment. The differences with the table 2 are in bold.

a close meaning and their characterization of the subject is the same. In the same way, the words "were" and "was" are close enough to have a low distance. In the alignment proposed in table 3, the alignment changed and we got a lower score.

Thanks to the adding of word embeddings, we can now detect and use the near matches in *WER*.

4. Dataset and baseline ASR, MT, SLT systems

For the experiments of this paper, we have extended our corpus presented in [20]. This corpus, available on a *github* repository[2] 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 will be updated soon 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.

LB:the part below on ASR and MT should be updated and modified by Benjamin and Tien

To obtain the speech transcripts (f_{hyp}), we built an ASR system based on KALDI toolkit [21]. The 3-gram language model was trained on the French ESTER corpus as well as French Gigaword (vocabulary size is 55k). CD-DNN-HMM based acoustic models were trained using the same ESTER corpus – see details in [22].

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 22.0% WER on *dev* set and 17.3% WER on *test* set. 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 translations (e_{hyp}), we used a French-English phrase-based translation system based on *moses* toolkit [23]. This medium-sized system was trained on Europarl and News parallel corpora (**mettre jour avec l'aide de Zied - qui a fourni le corpus - si besoin**.

Table 4 summarizes baseline ASR and MT performances obtained on our corpus.

task	ASR WER	MT BLEU (METEOR)	SLT BLEU (METEOR)
dev	22.00%	41.29% (40.19%)	30.57% (33.88%)
test	17.32%	51.20% (45.52%)	40.33% (40.10%)
overall	19.08%	47.45% (43.42%)	36.61% (37.68%)

Table 4: Baseline ASR, MT and SLT performance on our *dev* and *test* sets - BLEU and METEOR are scored w/o punctuation

5. Experiments and results

6. Discussion and analysis of translation exemples

7. Conclusions

8. Acknowlegements

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

9. References

[1] P. R. Dixon, A. Finch, C. Hori, and H. Kashioka, "Investigation on the effects of ASR tuning on speech translation performance," in *The proceedings of the International Workshop on Spoken Language Translation (IWSLT 2011)*, San Francisco, 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*. IEEE Signal Processing Society, Dec. 2015.

[4] 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, pp. 5632–5635.

[5] 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.

[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, Georgia: Association for Computational Linguistics, June 2013, pp. 746–751.

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

Filtering	dev			test			dev+test		
	WER	WER-E	WER-S	WER	WER-E	WER-S	WER	WER-E	WER-S
1best WER	22.00	18.09	17.38	17.32	12.94	12.35	19.08	14.88	14.25
best WER	11.99	10.35	10.09	7.23	5.67	5.47	9.02	7.43	7.21
best WER-E	12.14	9.89	9.71	7.38	5.26	5.12	9.17	7.00	6.85
best WER-S	12.15	9.96	9.66	7.37	5.30	5.09	9.17	7.06	6.81

Table 5: ASR scores obtained according to WER, WER-E and WER-S

Filtering	dev			test			dev+test		
	TER	BLEU	METEOR	TER	BLEU	METEOR	TER	BLEU	METEOR
1best	56.18	30.57	33.88	46.03	40.33	40.10	49.93	36.61	37.68
best WER	51.18	35.02	36.26	40.55	46.41	42.98	44.63	42.08	40.35
best WER-E	51.05	35.11	36.30	40.38	46.48	43.08	44.48	42.15	40.43
best WER-S	51.03	35.14	36.31	40.38	46.50	43.09	44.47	42.18	40.43

Table 6: SLT scores obtained against Post-Edition

[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, Arizona, USA, May 2013.

[9] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *The Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[10] 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, pp. 3111–3119.

[11] 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: Association for Computational Linguistics, October 2014, pp. 1724–1734.

[12] 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*. ACM, 2014, pp. 101–110.

[13] Y. Belinkov, M. Mohtarami, S. Cyphers, and J. Glass, "VectorSLU: A Continuous Word Vector Approach to Answer Selection in Community Question Answering Systems," in *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval*, vol. 15, 2015.

[14] 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.

[15] 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*, Lisbonne, Portugal, September 2015.

[16] 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.

[17] W. Y. Zou, R. Socher, D. M. Cer, and C. D. Manning, "Bilingual word embeddings for phrase-based machine translation." in *EMNLP*, 2013, pp. 1393–1398.

[18] R. E. Banchs, L. F. D'Haro, and H. Li, "Adequacy–fluency metrics: Evaluating mt in the continuous space model framework," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 3, pp. 472–482, March 2015.

[19] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[20] 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.

[21] 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.

[22] 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.

[23] 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 Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, République Tchèque, Juin 2007.