Better Evaluation of ASR in Speech Translation Context Using Word Embeddings
Ngoc-Tien Le, Christophe Servan, Benjamin Lecouteux, Laurent Besacier

To cite this version:

HAL Id: hal-01350102
https://hal.archives-ouvertes.fr/hal-01350102
Submitted on 29 Jul 2016

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.
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 reproducible 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 questioned. 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 substitution – 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 reproducible 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.

![Alignment chart](alignment_chart.png)

**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 example
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.

![Alignment chart](alignment_chart_2.png)

**Table 2:** WER-E estimation with word embeddings. Substitution score is replaced by a cosine distance, without questioning 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]).

\[
D_c(W_1, W_2) = 1 - S_c(W_1, W_2) 
\]

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.

![Alignment chart](alignment_chart_3.png)

**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 \( (f_{hyp}) \), verbatin transcript \( (f_{ref}) \), English text translation output \( (e_{hyp,nat}) \), speech translation output \( (e_{hyp,mt}) \), 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 referred to as dev set in the rest of the paper while the recently recorded part (4050 utterances) will be referred 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 transcriptions \( (f_{hyp}) \), 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/
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 \texttt{f}_{ref} reference is 21.92% WER on dev set and 17.46% WER on test set. This ASR system 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 \texttt{moses} phrase-based translation toolkit [19] to translate French ASR into English (\texttt{e}_{had}). 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 (\texttt{e}_{ref}).

### 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 3.4.


### 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 terms 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, MEETOR). Since BLEU and METEOR are not very efficient to evaluate translations at the sentence level, we prefer to use a more general measure like Pearson correlation.

---

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

<table>
<thead>
<tr>
<th>Tasks</th>
<th>metrics</th>
<th>ASR Ref.</th>
<th>ASR 1-best</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>WER</td>
<td>21.92</td>
<td>21.92</td>
</tr>
<tr>
<td></td>
<td>TER</td>
<td>38.84</td>
<td>55.64</td>
</tr>
<tr>
<td></td>
<td>BLEU</td>
<td>43.05</td>
<td>30.81</td>
</tr>
<tr>
<td></td>
<td>METEOR</td>
<td>40.73</td>
<td>34.02</td>
</tr>
<tr>
<td>test</td>
<td>WER</td>
<td>17.46</td>
<td>17.46</td>
</tr>
<tr>
<td></td>
<td>TER</td>
<td>45.64</td>
<td>58.70</td>
</tr>
<tr>
<td></td>
<td>BLEU</td>
<td>44.71</td>
<td>34.27</td>
</tr>
<tr>
<td></td>
<td>METEOR</td>
<td>39.10</td>
<td>34.27</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Tasks</th>
<th>metrics</th>
<th>ASR 1-best</th>
<th>Oracle from N-best</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>WER</td>
<td>21.92</td>
<td>12.01</td>
</tr>
<tr>
<td></td>
<td>WER-E</td>
<td>18.10</td>
<td>10.45</td>
</tr>
<tr>
<td></td>
<td>WER-S</td>
<td>17.41</td>
<td>10.19</td>
</tr>
<tr>
<td>test</td>
<td>WER</td>
<td>17.46</td>
<td>7.38</td>
</tr>
<tr>
<td></td>
<td>WER-E</td>
<td>13.13</td>
<td>5.86</td>
</tr>
<tr>
<td></td>
<td>WER-S</td>
<td>12.53</td>
<td>5.65</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>metrics</th>
<th>ASR 1-best</th>
<th>Oracle from N-best</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>TER</td>
<td>55.64</td>
<td>50.62</td>
</tr>
<tr>
<td></td>
<td>BLEU</td>
<td>30.81</td>
<td>35.29</td>
</tr>
<tr>
<td></td>
<td>METEOR</td>
<td>34.02</td>
<td>36.37</td>
</tr>
<tr>
<td>test</td>
<td>TER</td>
<td>54.70</td>
<td>54.13</td>
</tr>
<tr>
<td></td>
<td>BLEU</td>
<td>34.27</td>
<td>39.34</td>
</tr>
<tr>
<td></td>
<td>METEOR</td>
<td>34.27</td>
<td>36.55</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Comparison</th>
<th>TER</th>
<th>BLEU</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>O. WER-E best</td>
<td>255</td>
<td>310</td>
<td>321</td>
</tr>
<tr>
<td></td>
<td>O. WER best</td>
<td>190</td>
<td>271</td>
<td>315</td>
</tr>
<tr>
<td></td>
<td>Ties</td>
<td>2198</td>
<td>2062</td>
<td>2007</td>
</tr>
<tr>
<td>Test</td>
<td>O. WER-E best</td>
<td>341</td>
<td>451</td>
<td>510</td>
</tr>
<tr>
<td></td>
<td>O. WER best</td>
<td>264</td>
<td>381</td>
<td>399</td>
</tr>
<tr>
<td></td>
<td>Ties</td>
<td>3445</td>
<td>3218</td>
<td>3141</td>
</tr>
</tbody>
</table>

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.

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) 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.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>metrics</th>
<th>ASR optimized with WER-E</th>
<th>ASR optimized with WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>TER</td>
<td>55.64</td>
<td>55.52</td>
</tr>
<tr>
<td></td>
<td>BLEU</td>
<td>30.81</td>
<td>30.84</td>
</tr>
<tr>
<td></td>
<td>METEOR</td>
<td>34.02</td>
<td>34.00</td>
</tr>
<tr>
<td>test</td>
<td>TER</td>
<td>58.71</td>
<td>58.56</td>
</tr>
<tr>
<td></td>
<td>BLEU</td>
<td>34.27</td>
<td>34.38</td>
</tr>
<tr>
<td></td>
<td>METEOR</td>
<td>34.27</td>
<td>34.26</td>
</tr>
</tbody>
</table>

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

5.5. Translation examples
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 modified 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 reproducible experiments, code allowing to call our modified WER is made available on github\(^\text{3}\).

7. Acknowledgements
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


