Modifications of Machine Translation Evaluation Metrics by Using Word Embeddings

Haozhou Wang
Department of Linguistics
University of Geneva
Haozhou.Wang@etu.unige.ch

Paola Merlo
Department of Linguistics
University of Geneva
Paola.Merlo@unige.ch

Abstract

Traditional machine translation evaluation metrics such as BLEU and WER have been widely used, but these metrics have poor correlations with human judgements because they badly represent word similarity and impose strict identity matching. In this paper, we propose some modifications to the traditional measures based on word embeddings for these two metrics. The evaluation results show that our modifications significantly improve their correlation with human judgements.

1 Introduction

One of the challenges for Machine Translation (MT) research is how to evaluate the quality of translations automatically and correctly. Earlier word-based metrics such as BLEU (Papineni et al., 2002), WER and TER (Snover et al., 2006) have been widely used in machine translation, but these metrics have poor correlations with human judgements, especially at the sentence level. One reason is that they just allow strict string matchings between hypothesis and references. For example, the semantically related words "learn" and "study" and words that differ only by morphological markers, such as "study" and "studies" are considered different words although they have a similar meaning. The traditional solution for improving their performance is to use more references. However, multiple references are rare and expensive. Moreover, these n-gram-based evaluations have been shown to be biased in favour of statistical methods, largely because they do not allow grammatically-costrained lexical freedom.

In recent years, many proposals have been put forth and new metrics have appeared and shown their good performance (Machacek and Bojar, 2013; Machacek and Bojar, 2014; Stanojević et al., 2015). However, improving the performance of existing metrics does not require developing a whole new metric. Proposals that modify existing metrics and show competitive results have also been proposed. One of the common solutions to improve traditional metrics consists in changing strict string matching to fuzzy matching at the surface level. For example, LeBLEU (Virpioja and Grönroos, 2015) — a variant of standard BLEU, also called "Letter-edit-BLEU" or "Levenshtein-BLEU" — takes into account letter-edit distance — Levenshtein distance including the spaces between the words — between hypothesis and references instead of strict n-gram matchings. More recently, Weiyue et al. (2016) have proposed a character-level TER (CharacTER) which calculates the character-level edit distance, while still performing the shift edits at the word level. The evaluation results show that this kind of modifications have a good effect on string-level similar words, but that they don't work well on words that are semantically similar, but are orthografically different strings.

To capture semantic similarity, one established way is to apply additional linguistic knowledge, such as synonym dictionaries. For example, TER-Plus (Snover et al., 2009) use WordNet (Fellbaum, 1998) to compute synonym matches in addition to the four original operations (Insertion, Deletion, Substitution and Shift). Although such linguistic resources are helpful, they are often lacking in coverage and affect computation speed and ease of use.

Current research on word embeddings (Bengio et al., 2003; Mikolov et al., 2013) maps each word to a low-dimensional vector. The vectors of the words that are semantically similar have been shown to be close to each other in vector space. The similarity between words then can be captured by calculating

Proceedings of the Sixth Workshop on Hybrid Approaches to Translation,
pages 33–41, Osaka, Japan, December 11, 2016.

the geometric distance between their vectors. On this basis, Le and Mikolov (2014) extend word-level representation to sentence and document level, which allows them to compute the similarity between two sequence of words. Recently, this kind of vector representation has been widely integrated in MT evaluation. Banchs et al. (2015) use Latent Semantic Indexing to project sentences as bag-of-words into a low-dimensional continuous space to measure the adequacy on an hypothesis. A monolingual continuous space has been used to capture the similarity between hypothesis and reference and a cross-language continuous space has been used to calculate the similarity between source sentence and hypothesis. With the same idea, Vela and Tan (2015) proposed a Bayesian Ridge Regressor which use document-level embeddings as features and METEOR score as target to predict the adequacy of hypothesis. The study of Chen and Guo (2015) uses vector representation more directly. In their study, each sentence has been transformed into a vector (they tried 3 kinds of vector representation: one-hot, word embedding and recursive auto-encoder representations). The evaluation score is calculated by the distance between the hypothesis vector and the reference vector, with a length penalty. More recently, Servan et al. (2016) combine word embeddings and DBnary (Sérasset, 2015), a multilingual lexical resource, to enrich METEOR.

In this paper, we also incorporate word embeddings in our similarity score to improve machine translation evaluation metrics. We propose measures that, while being largely compatible with previous proposals (BLEU and WER), include semantic word similarity and improve on the state of the art. Differently from with the above-mentioned works, our approach simply uses monolingual word embeddings, and still has competitive performance at both sentence and system level.

Because these measures are modifications of BLEU and WER (we call them BLEU$_{modif}$ and WER$_{modif}$), they also support systematic comparisons of results: if BLEU$_{modif}$ or WER$_{modif}$ is better correlated with human judgments because word embeddings allow it to better captures lexical semantic similarity, then the improvement in performance must be due to the fact that the system translation exhibits lexical semantic variation. These modified measures then allow us to compare different architectures according to their amount of lexical variation. Compared to the standard BLEU and WER versions, which have been argued to penalize rule-based systems more, these modified measures do not penalize systems based on their architecture. This gives us the possibility to evaluate fairly both the rule-based and the statistical components of a hybrid system.

In this paper, we will first descible our method in next section. Our experimental results in section 3 show that even a simple modification could significantly improve the performances over traditional metrics.

2 Method

The standard BLEU and WER metrics compute strict matching between n-grams or words. Our modifications for these two metrics is to use a similarity score between n-grams (words for WER) instead of strict matching. It has previously been shown that word embeddings represent the contextualised lexical semantics of words (Mikolov et al., 2013; Bengio et al., 2003). We first use the popular toolkit Word2Vec[1] provided by Mikolov et al. (2013) to train our word embeddings. At the word level, the similarity score between two words is the cosine similarity between word vectors. At the n-gram level, we average the vectors of all words in the n-gram and use the similarity between average vectors as the n-gram similarity score. All the Out-Of-Vocabulary words are skipped when computing the similarity score. For example, word vectors show that "study" and "studies" are very similar, while "study" and "play" are not very similar.

- Vector of "study" is [0.1049, -0.1103, ..., 0.0752]
- Vector of "studies" is [0.0035, -0.0799, ..., 0.1178]
- Vector of "play" is [-0.0250, 0.0531 ..., 0.0759]
- Similarity score of "study" and "studies": 0.534

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

- Similarity score of "study" and "play": 0.058

Word2Vec provides two embedding algorithms, Skip-Gram and Continuous Bag-of-Words (CBOW). The study of Levy et al. (2015) and Mikolov et al. (2013) show that Skip-Gram better represents word similarity, but Baroni et al. (2014) show the opposite. In our study, we will use both of them, and try to find the better one for our modifications of BLEU and WER.

Our Python program uses the Gensim package[2] for implementing the trained word embeddings. The code of our modified measures is provided on the Github page[3] .

2.1 Modification for BLEU metric

The original BLEU score is calculated with the modified n-gram precision P_n and the brevity penalty BP, as shown in (1).

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n log P_n) \tag{1}$$

where w_n is a positive weight which is used to adjust the proportions of different n-grams. In the baseline of Papineni et al. (2002), they used N = 4 and uniform w_n = 1/4. The brevity penalty BP is used to penalise the translations that are shorter than their references.

The modified n-gram precision P_n is the proportion of matched n-grams ng between the translation sentence T and the corresponding reference sentence, shown in (2) and (3).

$$P_n = \frac{\sum_{ng \in T} C_{clip}(ng)}{\sum_{ng \in T} C(ng)} \tag{2}$$

$$C_{clip}(ng) = min\{C(ng), MaxC_{ref}(ng)\} \tag{3}$$

Here, $C_{clip}(ng)$ is called clipped counts, $MaxC_{ref}(ng)$ is the maximum value of the corresponding of matched n-gram in the reference.

One of BLEU's disadvantages is that the precision P_n considers a valid match only for those words that are identical between translations and references. We propose a modification for BLEU that instead of using the modified n-gram precision P_n uses the similarity n-gram precision P_{sim}, which is defined in (4).

$$P_{sim} = \frac{\sum_{ng \in T} Max_{sim_{pruned}}(ng, T, R, \gamma)}{\sum_{ng \in T} C(ng)} \tag{4}$$

In this formula, P_{sim} is computed as follows:

- Calculate the similarity scores between an n-gram(ng) in the translation sentence T and all the n-grams in the reference sentence R.
- Prune the maximum similarity score with a threshold γ.
- Sum the $Max_{sim_{pruned}}(ng, T, R, \gamma)$ of all the n-grams in T and divide the result by the the number of n-grams in T.

Our modified BLEU metric is defined in (5).

$$BLEU_{modif} = BP \cdot exp(\sum_{n=1}^{N} w_n log P_{sim}) \tag{5}$$

Same as Papineni et al. (2002), in our baseline, we use N = 4 and uniform w_n = 1/4. We will tune the threshold γ and try to find the best threshold.

[2]http://radimrehurek.com/gensim/index.html
[3]https://github.com/ChatonPatron/VecEval

2.2 Modification for WER metric

The standard word error rate is computed in the following way:

$$WER = \frac{S + D + I}{N} \tag{6}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the number of words in the reference sentence. Normally, every action has a weight of 1, whether it is substitution, deletion or insertion. Assigning equal weights to all actions does not represent the intuition that the cost of a substitution depends on the similarity of the words. For example, if the cost of the needed operations is a measure of how hard it is to recover the real translation from the system translation, then the effort is not always the same, it depends on the quality of the translation. For example, if the translation word simply has a morphological error, the action "substitution" will be very easy, but if the translation word is completely different from the correct word, this action will be definitely harder. Our modification for the WER metric (WER_{modif}) focusses on the action "substitution": instead of giving the same weight to the three operations, we calculate their weights as shown in (7).

$$S_{modif} = 1 - Score_{sim}(word_{old}, word_{new}); D, I = 1 \tag{7}$$

Here, $Score_{sim}(word_{old}, word_{new})$ is the similarity score between the old word and the substituted word. Same with the standard WER, a higher score means a worse translation.

3 Experiments

We carried out some experiments to study our modified metrics. The experiments are based on the English-to-French, English-to-German and French-to-English, German-to-English data provided for the metrics task of the Workshops on Statistical Machine Translation (WMT) (Stanojević et al., 2015; Machacek and Bojar, 2014). This kind of data consists of human judgements for the outputs of different MT systems. The principle of the experiments is to tune and evaluate our modified metrics by measuring the correlation between our scores and the human judgement scores at the segment-level and at the system-level. The segment-level correlation is calculated by the Kendall's rank correlation coefficient and the system-level correlation is calculated by Pearson's correlation coefficient. We use the dataset of WMT-14[4] for the tuning task and WMT-15[5] for the evaluation task.

Our word embedding models are trained on a multilingual corpus called "News Crawl" shared by WMT-16[6]. This corpus contains a large amount of news articles from 2007 to 2015 in different languages. The size of our training data is 2.917 billion words for English, 0.877 billion words for French and 1.752 billion words for German. For each language, we trained two embedding models with the two different algorithms Skip-Gram (Vector Size = 500, Window Size = 10) and CBOW (Vector Size = 500, Window Size = 5)

3.1 Parameter Tuning

We first ran a grid search of ten values to tune the parameter γ (from $\gamma = 0.0$ to $\gamma = 0.9$) on the dataset of WMT-14. The results are reported in Figure 1. If we look at the figure of Skip-Grams (left), we find that the curves at the segment-level are very similar, the correlation score improves after $\gamma = 0.3$, but reduces quickly after $\gamma = 0.7$. The curves at the system level are quite different. For French,German-to-English, the correlation score gets a little improvement after $\gamma = 0.3$, but for English-to-French,German, the correlation score decreases directly after $\gamma = 0.3$. For the figure of CBOW, the curves are very similar. The correlation score stabilizes before $\gamma = 0.3$, and decreases after. Differently from Skip-gram, the correlation at the segment-level drops more quickly than the correlation at the system-level.

The tuning results reported in Table 1 give the numerical values of the best correlation scores. We can conclude that, for the modified BLEU measure (BLEU_{modif}), the best result at the segment-level (two

[4]http://www.statmt.org/wmt14/metrics-task/

[5]http://www.statmt.org/wmt15/metrics-task/

[6]http://www.statmt.org/wmt16/translation-task.html

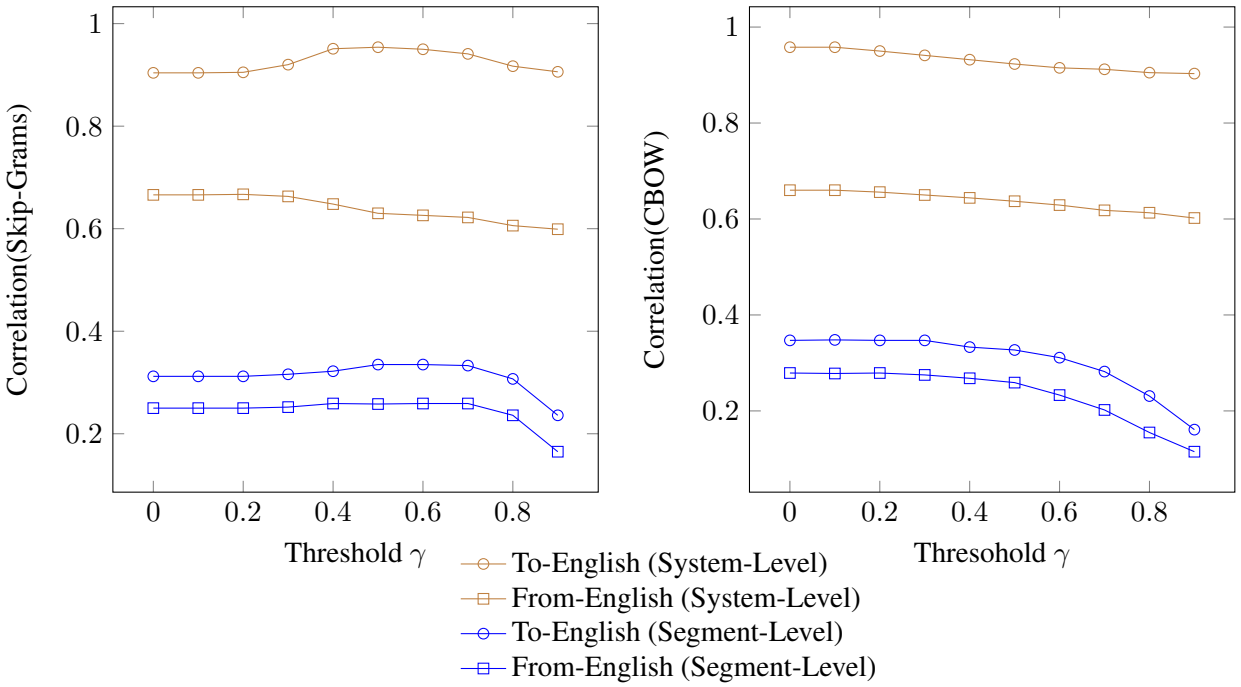

Figure 1: Results of tuning the threshold γ for modified BLEU on the WMT-14 dataset. To-En includes French and German to English. From-En includes English to French and German. The correlation scores are the averages of the two source (or target) languages. Note that in the WMT-14 metric task, all the results into German at the system-level are markedly lower than the others.

directions) and French,German-to-English at the system-level are produced by the CBOW algorithm with a threshold equal to 0.1. Skip-Gram with a threshold equal to 0.2 works best for the English-to-French,German system-level measure. For modified WER (WER_{modif}), CBOW always has a better result than Skip-Gram.

		Segment-Level				System-Level			
		To-En		From-En		To-En		From-En	
		Corr.	γ	Corr.	γ	Corr.	γ	Corr.	γ
BLEU_{modif}	Skip-Gram	0.335	0.6	0.259	0.5	0.954	0.5	**0.667**	0.2
	CBOW	**0.348**	0.1	**0.278**	0.1	**0.957**	0.1	0.660	0.1
WER_{modif}	Skip-Gram	0.332	-	0.253	-	0.942	-	0.662	-
	CBOW	**0.351**	-	**0.277**	-	**0.956**	-	**0.671**	-

Table 1: Tuning results: The results for modified BLEU shown in this table are the results of different embedding algorithms with the best threshold γ. To-En includes French and German to English. From-En includes English to French and German. The correlation scores are the averages of the languages mentioned.

3.2 Performance Evaluation

We evaluated our modified metrics on the dataset of WMT-15 with the best parameters found in the tuning phase. For a better understanding of the general performance of our measures, we compared our modified metrics with standard BLEU, sentence-level smoothed BLEU, TER, NIST and WER. The results reported in Table 2 show that, compared with their original versions, both the modified BLEU or the modified WER show an improvement on the correlation with human judgements, both at the segment-level and at the system-level. Their performance is much better than TER and NIST (Doddington, 2002), especially on the English-to-French,German data. If we observe the ranking of metrics, we find that

	Segment-Level				System-Level			
	To-En		From-En		To-En		From-En	
	Corr.	Rank	Corr.	Rank	Corr.	Rank	Corr.	Rank
Top	0.438	1/22	0.373	1/15	0.984	1/25	**0.922**	**1/18**
TER	-	-	-	-	0.935	20/25	0.756	13/18
NIST	-	-	-	-	0.941	17/25	0.726	15/18
BLEU	0.137	22/22	0.139	15/15	0.920	22/25	0.760	12/18
Sent-BLEU	0.359	19/22	0.306	13/15	-	-	-	-
BLEU$_{modif}$	**0.390**	14/22	**0.353**	7/15	**0.951**	13/25	**0.881**	7/18
WER	**0.373**	17/22	0.324	12/15	0.930	21/25	0.754	14/18
WER$_{modif}$	**0.397**	11/22	0.347	8/15	**0.949**	15/25	**0.922**	1/18

Table 2: System-level and segment-level correlation with the human judgement on the WMT-15 dataset. To-En includes French and German to English. From-En includes English to French and German. The correlation scores are the averages of the languages mentioned.

after our modifications, the ranks of BLEU and WER are increased by at least four or five ranks. For English-to-French,German system-level, the modified WER becomes the top metric among eighteenth participants.

The results show that a measure that simply augments matching by a similarity notion has better performance than strict string matching, and that current word embeddings techniques capture this notion of similarity.

	BLEU	BLEU$_{modif}$	WER	WER$_{modif}$
Hyp1:	0.508	0.835	0.333	0.178
Hyp2:	0.508	0.812	0.333	0.199
Hyp3:	0.508	0.797	0.333	0.219

Table 3: Single translation evaluation scores.

A qualitative analysis of results also shows that the captured notion of similarity corresponds to ranking of sentence alternatives by native speakers. For example, looking at some randomly chosen individual sentences, we find some interesting examples: The source sentence “History is a great teacher” is translated as “Die Geschichte ist ein großartiger Lehrmeister” in German. The following hypotheses are the output translations of three MT systems from WMT-15 translation task.

- Hypothesis 1: Die Geschichte ist ein guter Lehrer.
- Hypothesis 2: Die Geschichte ist ein großer Lehrer.
- Hypothesis 3: Die Geschichte ist ein großer Meister.

We used the original BLEU and WER and our modified versions to evaluate these three hypotheses. The scores are shown in Table 3. Before our modifications, the original BLEU and WER metrics give the same scores to these three different hypotheses. After our modifications, the modified measures are able to recognize the difference. According to a native German speaker, the rank of these hypotheses is : Hyp1>Hyp2>Hyp3. This rank is the same as what is proposed by the modified measures, showing that the measure is not only more accurate within a system, but also more sensitive to differences across systems.

When we observe the system-level scores of different participants of WMT-15 Translation Task, we find an interesting phenomenon. According to the human evaluation scores, for the English-to-German systems, the only Rule-based system “PROM-RULE” is ranked third among sixteen MT systems. The score of an online system “Online-A” is slightly lower but very close. According to the official report of

the WMT-15 Translation Task (Bojar et al., 2015), these two systems are considered tied. However, if we re-rank all the systems by standard BLEU or WER, according to the results reported in Figure 2 and Table 4, we find that the rank of "PROM-RULE" decreases quickly from number three to number ten or eleven, and the rank of "Online-A" becomes much higher than "PROM-RULE". It is in fact well-known that because rule-based systems usually apply some dictionary resources, their lexical variation is richer than other kinds of MT systems. But this is the reason why these kinds of systems are usually considered good according to human judgements, but not as good when scored automatically. Our modifications changed the situation: we give the rule-based system the opportunity to score correctly by similar words. So that the rank of our modified metrics is similar to the rank of the human evaluation. Note that, for the modified BLEU, the scores are very close (the difference between the scores is less than 0.001), so that we can consider that, like the human judgements, they are at the same level.

	Human	BLEU	BLEU_{modif}	WER	WER_{modif}
PROM-RULE	0.2600	0.2253	0.7297	0.6887	0.5866
Online-A	0.2350	0.1859	0.7302	0.6284	0.5810

Table 4: English-to-German system-level evaluation scores of "PROM-RULE and "Online-A" (Systems from WMT-15 Translation Task)

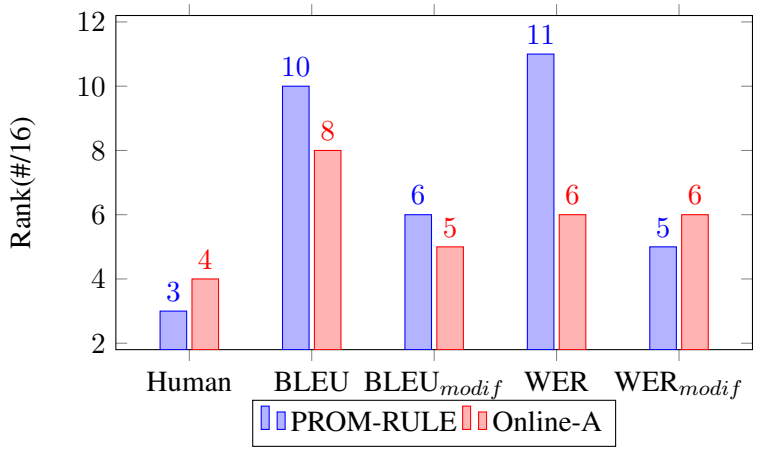

Figure 2: English-to-German system-level ranking of "PROM-RULE and "Online-A" (Systems from WMT-15 Translation Task)

4 Conclusions

In this paper, we have described our modifications for BLEU and WER metrics based on word-embeddings. The modifications allow these measures to take into account the semantic similarity of the words or of the n-grams, and not just string similarity. With this kind of semantic similarity, BLEU and WER do not penalize rule-based systems or rule-based components of hybrid systems more than statistical systems and lead to a fairer evaluation. Experiments on the WMT-15 metric task dataset shows that, compared to the standard BLEU and WER, the modified metrics obtains a better correlations with human judgments both at the segment-level and at the system-level. The improvement is quite apparent for the English-to-French,German data.

References

Rafael E Banchs, Luis F DHaro, and Haizhou Li. 2015. Adequacy–fluency metrics: Evaluating MT in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247, Baltimore, June.

Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of machine learning research*, 3:1137–1155.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September.

Boxing Chen and Hongyu Guo. 2015. Representation-based Translation Evaluation Metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 150–155, Beijing, July.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, California, March.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, May.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1188–1196, Beijing, June.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, May.

Matouvs Machacek and Ondrej Bojar. 2013. Results of the WMT13 Metrics Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, August.

Matouvs Machacek and Ondrej Bojar. 2014. Results of the WMT14 Metrics Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, June.

Tomas Mikolov, Ilya Sutskever, Chen Kai, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in neural information processing systems*, pages 3111–3119.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Fortieth Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, July.

Gilles Sérasset. 2015. DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF. *Semantic Web*, 6(4):355–361.

Christophe Servan, Alexandre Bérard, Zied Elloumi, Hervé Blanchon, and Laurent Besacier. 2016. Word2Vec vs DBnary: Augmenting METEOR using Vector Representations or Lexical Resources? *arXiv preprint arXiv:1610.01291*.

Matthew G Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, volume 200, Boston, August.

Matthew G Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Ter-plus: Paraphrase, semantic, and alignment enhancements to translation edit rate . *Machine Translation*, 23:117–127.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondrej Bojar. 2015. Results of the WMT15 Metrics Shared Task . In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, September.

Mihaela Vela and Liling Tan. 2015. Predicting Machine Translation Adequacy with Document Embeddings. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 402–410, Lisbon, September.

Sami Virpioja and Stig-Arne Grönroos. 2015. LeBLEU: N-gram-based Translation Evaluation Score for Morphologically Complex Languages. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 411–416, Lisbon, September.

Wang Weiyue, Peter Jan-Thorsten, Rosendahl Hendrik, and Ney Hermann. 2016. CharacTER: Translation Edit Rate on Character Level. In *Proceedings of the First Conference on Machine Translation*, pages 505–510, Berlin, August.