Predicting Machine Translation Adequacy with Document Embeddings

Mihaela Vela
Saarland University
Saarbrücken, Germany
m.vela@mx.uni-saarland.de

Liling Tan
Saarland University
Saarbrücken, Germany
liling.tan@uni-saarland.de

Abstract

This paper describes USAAR's submission to the the *metrics shared task* of the Workshop on Statistical Machine Translation (WMT) in 2015. The goal of our submission is to take advantage of the semantic overlap between hypothesis and reference translation for predicting MT output adequacy using language independent document embeddings. The approach presented here is learning a Bayesian Ridge Regressor using document skip-gram embeddings in order to automatically evaluate Machine Translation (MT) output by predicting semantic adequacy scores. The evaluation of our submission – measured by the correlation with human judgements – shows promising results on system-level scores.

1 Introduction

Translation is becoming an utility in everyday life. The increased availability of real-time machine translation services relying on Statistical Machine Translation (SMT) allows users who do not understand the language of the source text to quickly gist the text and understand its general meaning. For these users, accurate meaning of translated words is more important than the fluency of the translated sentence.

However, SMT suffers from poor lexical choices. Fluent but inadequate translations are commonly produced due to the strong bias towards the language model component that prefers consecutive words based on the data that the system is trained on.

Current state of art MT evaluation metrics are generally able to identify problems with grammaticality of the translation but less evidently accuracy of translated semantics, e.g. incorrect translation of ambiguous words or wrong assignment of semantic roles. In the example below, the ideal Machine Translation (MT) evaluation metric should appropriately penalise poor lexical choice, such as *braked*, and reward or at least allow leeway for semantically similar translations, such as *external trade*.

Source (DE):
Auch der Auenhandel bremste die Konjunktur.

Phrase-based MT:
The foreign trade braked the economy.

Neural MT:
External trade also slowed the economy.

Reference (EN):
Foreign goods trade had slowed, too.

The German word *bremste* is commonly used as *braked* in the context of driving, but the appropriate translation should have been *slowed* in the example mentioned above. Although the phrase *external trade* differs from *foreign goods trade* in the reference sentence, it should be considered as an acceptable translation.

We propose a semantically grounded, language independent approach using Semantic Textual Similarity (STS) to evaluate the adequacy of the machine translation outputs with respect to their reference translations.

The remainder of this paper is structured as follows. Section 2 gives an overview of the related work in the field of MT evaluation. Section 3 presents the approach behind the USAAR submission to the metrics shared task. In Section 4 we present the data and experiments for this submission. Section 5 covers the evaluation of our metric by the WMT2015 metrics task organisers and in Section 6 we conclude on our WMT2015 metrics task submission.

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 402–410,
Lisboa, Portugal, 17-18 September 2015.

2 Related Work

Researchers in the field of MT evaluation have proposed a large variety of methods for assessing the quality of automatically produced translations. Approaches range from fully automatic quality scoring to efforts aimed at the development of "human" evaluation scores that try to exploit the (often tacit) linguistic knowledge of human evaluators.

2.1 Automatic Evaluation of MT

MT output is usually evaluated by automatic language-independent metrics that can be applied to MT output, independent of the target language. Automatic metrics typically compute the closeness (adequacy) of a hypothesis to a reference translation and differ from each other by how this closeness is measured. The most popular MT evaluation metrics are IBM BLEU (Papineni et al., 2002) and NIST (Doddington, 2002), used not only for tuning MT systems, but also as evaluation metrics for translation shared tasks, such as the Workshop on Statistical Machine Translation (WMT).

IBM BLEU uses n-gram precision by matching machine translation output against one or more reference translations. It accounts for adequacy and fluency by calculating word precision, i.e. the n-gram precision.

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (1)$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ -e^{(1-r/c)} & \text{if } c \leq r \end{cases} \quad (2)$$

In order to deal with the over generation of common words, precision counts are clipped, meaning that a reference word is exhausted after it is matched against the same word in the hypothesis. This is then called the modified n-gram precision. For BLEU, the modified n-gram precision is calculated with N=4, the results being combined by using the geometric mean. Instead of recall, BLEU computes the Brevity Penalty (BP) (see formula in 2), thus penalising candidate translations which are shorter than the reference translations.

The NIST metric is derived from IBM BLEU. The NIST score is the arithmetic mean of modified n-gram precision for N=5 scaled by the BP. Additionally, NIST also considers the information gain of each n-gram, giving more weight to more informative (less frequent) n-grams and less weight to less informative (more frequent) n-grams.

Another often used machine translation evaluation metric is METEOR (Denkowski and Lavie, 2014). Unlike IBM BLEU and NIST, METEOR evaluates a candidate translation by calculating precision and recall on the unigram level and combining them into a parametrised harmonic mean. The result from the harmonic mean is then scaled by a fragmentation penalty which penalizes gaps and differences in word order. METEOR is described in detail in Section 3.1.

Besides these evaluation metrics, several other metrics are used for the evaluation of MT output. Some of these are the WER (word error-rate) metric based on the Levensthein distance (Levenshtein, 1966), the position-independent error rate metric PER (Tillmann et al., 1997) and the translation edit rate metric TER (Snover et al., 2006) with its newer version TERp (Snover et al., 2009).

The semantics of both hypotheses and reference translations is considered by MEANT (Lo et al., 2012). MEANT, based on HMEANT (Lo and Wu, 2011a; Lo and Wu, 2011b; Lo and Wu, 2011c), is a fully automatic semantic MT evaluation metric, measuring semantic fidelity by determining the degree of parallelism of verb frames and semantic roles between hypothesis and reference translations. Some approaches aim at combining several linguistic and semantic aspects. Gonzàlez et al. (2014) as well as Comelles and Atserias (2014) introduce their fully automatic approaches to machine translation evaluation using lexical, syntactic and semantic information when comparing the machine translation output with reference translations.

2.2 Human Evaluation of MT

Human MT evaluation approaches employ the knowledge of human annotators to assess the quality of automatically produced translations along the two axes of target language correctness and semantic fidelity. The Linguistics Data Consortium (LDC) introduced a MT evaluation task that elicits quality judgement of MT output from human annotators using a numerical scale (Linguistics Data Consortium, 2005). These judgements were split into two categories: adequacy, the degree of meaning preservation, and fluency, target language correctness.

Adequacy judgements require annotators to rate the amount of meaning expressed in the reference translation that is also present in the translation hypothesis. Fluency judgements require annotators to rate how well the translation hypothesis in the target language is formed, disregarding the sentence meaning. Although evaluators are asked to assess the fluency and adequacy of a hypothesis translation on a Likert scale separately, Callison-Burch et al. (2007) reported high correlation between annotators' adequacy and fluency scores.

MT output is also evaluated by measuring human post-editing time for productivity (Guerberof, 2009; Zampieri and Vela, 2014), or by asking evaluators to rank MT system outputs (by ordering a set of translation hypotheses according to their quality). Vela and van Genabith (2015) show that this task is very easy to accomplish for evaluators, since it does not imply specific skills, a homogeneous group being enough to perform this task. This is also the method applied during the last years WMTs, where humans are asked to rank machine translation output by using APPRAISE (Federmann, 2012), a software tool that integrates facilities for such a ranking task.

An indirect human evaluation method, that is also employed for error analysis, are reading comprehension tests (e.g. Maney et al. (2012), Weiss and Ahrenberg (2012)). Other evaluation metrics try to measure the effort that is necessary for "repairing" MT output, that is, for transforming it into a linguistically correct and faithful translation. One such metric is HTER (Snover et al., 2006), which uses human annotators to generate targeted reference translations by means of post-editing, the rationale being that by this the shortest path between a hypothesis and its correct version can be found.

2.3 Semantic Textual Similarity

Given two snippets of text, the Semantic Textual Similarity (STS) task attempts to measure their semantic equivalence on a scale of 1 to 5 (Agirre et al., 2014). The STS task is organized annually during the SemEval workshop and systems are evaluated based on their Pearson correlation coefficient with the human annotations.

The STS is similar to the task of determining the adequacy of a translation hypothesis with respect to a reference translation. The STS task is usually treated as a regression task where systems are trained using features such as:

(i) linguistics annotation overlaps between the two text snippets, e.g. syntactic dependency, lexical paraphrases, part of speech (Šarić et al., 2012; Han et al., 2012; Pilehvar et al., 2013)

(ii) machine translation metrics as features in training a supervised regressor (Rios et al., 2012; Barrón-Cedeño et al., 2013; Huang and Chang, 2014; Tan et al., 2015b)

(iii) word/document embeddings similarity (Sultan et al., 2015; Arora et al., 2015).

Linguistic annotations are restricted by the availability of the annotation tools, that are often language dependent. Machine translation evaluation metrics generally provide a shallow comparison between hypotheses and reference translations focusing on capturing the grammatical similarities between the texts, whereas the use of document embeddings focuses on capturing the semantic similarity between texts. Word embeddings dates back to the traditional Latent Semantic Analysis (LSA) vector spaces used for information retrieval (Landauer and Dutnais, 1997) to the current trend of using neural nets for NLP/MT tasks (Bordes et al., 2011; Huang et al., 2012; Bordes et al., 2012; Chen and Manning, 2014; Bowman et al., 2015).

3 Our Approach

Although consensus exists that lexical-based metrics cannot cover the entire range of linguistic phenomena (Vela et al., 2014a; Vela et al., 2014b), the goal in the MT community remains to have a language independent metric that takes into account for lexical, syntactic and semantic information when mapping the MT output against the reference translation. The questions that have to be accounted for in such a language-independent metric are:

(i) Is there a lexical overlap between reference and hypothesis translation?

(ii) Is there a syntactic overlap between reference and hypothesis translation?

(iii) Is there a semantic overlap between reference and hypothesis translation?

In the ideal situation one would also take into account lexical, syntactic and semantic information from the source text. Specific information (on lexical, syntactic, semantic level) from the source text could help improving not only the translation process, but also the evaluation.

As pointed out in Section 2, there are several approaches which tend to cover the entire range of linguistic phenomena in the evaluation process. The approach presented in this paper is leaned on the STS approach, mentioned in Section 2.3, aiming to provide a language independent adequacy score using document embedding similarity as opposed to the traditional synonyms and paraphrase overlap approach used in METEOR. The matching of synonyms in METEOR relies on WordNet (Miller, 1995), which is a limited resource, making it impossible to use the synonymy module from METEOR for other languages than English. The provided or self-extracted paraphrase tables for METEOR are available only for languages for which big corpora are available, making it difficult to provide paraphrases for under-resourced languages. Since METEOR relies on the WordNet synonymy and language dependent paraphrase tables for its semantic component, our goal is to substitute this components with a language independent component.

Different from the STS task, the WMT metrics task provides the ranks of the systems' hypotheses instead of absolute human evaluation scores of the translation hypotheses. To generate the absolute scores, we use the METEOR scores between the translation hypotheses and the reference translations.

To induce the word embeddings, we trained a skip-gram model phrasal word2vec neural net (Mikolov et al., 2013) using `gensim` (Řehůřek and Sojka, 2010). The neural nets were trained to produce 400 dense features for 100 epochs with a window size of 5 for all words from the WMT metrics task data.

$$v(doc) = \frac{\sum_i^n v(w_i)}{n} \qquad (3)$$
$$doc = \{w_1, ..., w_n\}$$

To generate the document embeddings, *v(doc)*, we sum the word embeddings from the document and normalised it by the number of words. The setup for the skip-gram model and the document vector is similar the techniques uses in STS tasks (Sultan et al., 2015; Tan et al., 2015a).

$$sim(hyp, ref) = v(hyp) \cdot v(ref) \qquad (4)$$

The document embedding similarity is achieved by the dot product between the translation hypothesis (*hyp*) and the reference translation (*ref*). Geometrically, the dot product between the hypothesis and the reference translation yields the cosine similarity between two vectors. Alternatively, one could also calculate the cosine similarity by summing the square of the word vector of the intersecting word embeddings and normalise the document by the root of the sum square for all words in the documents (Tan, 2013)[1].

Using the similarity scores between the hypothesis and reference embeddings, we train a Bayesian Ridge Regressor targeting the METEOR scores as the desired output.

3.1 METEOR

METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Denkowski and Lavie, 2014) is an MT evaluation metric which tries to consider both grammatical and semantic knowledge. The metric is based on the alignment between a hypothesis translation and a reference translation containing four modules. The number of modules to be used depends on the availability of resources for a specific language. The first module generates the alignments based on the surface forms of the words in the hypothesis and reference translation. The next module performs the alignment on word stems, followed by the alignment of words listed as synonyms in WordNet (Miller, 1995). The last module is responsible for the paraphrase matching between the hypothesis and reference translation, based on the provided or the self-extracted paraphrase tables. For the final score calculation all matches are generalised to phrase/chunk matches with a start position and phrase length in each sentence.

Different from other evaluation metrics, METEOR makes the distinction between content words and function words in the hypothesis (h_c, h_f) and reference (r_c, r_f) translation. This distinction is made by a provided function words list.

[1] An implementation of the alternative cosine can be fount at `http://tinyurl.com/pywsd-cosine`. The original implementation is reported in (Tan and Bond, 2013)

From the final alignment between hypothesis and reference translation, precision (P) and recall (R) is calculated by weighting content words and function words differently. This is described by Denkowski and Lavie (2014) as follows. For each of the matchers (m_i) count the number of content and function words covered by matches of this type in the hypothesis $(m_i(h_c), m_i(h_r))$ and reference $(m_i(r_c), m_i(r_r))$ translation. The weighted precision (P) and recall (R) is computed by using the matcher weights $w_i ... w_n$ and the function word weight γ as shown in 5 and 6.

$$P = \frac{\sum_i w_i \cdot (\delta \cdot m_i(h_c) + (1-\delta) \cdot m_i(h_c))}{\gamma \cdot \mid h_c \mid + (1-\gamma) \cdot \mid h_f \mid} \tag{5}$$

$$R = \frac{\sum_i w_i \cdot (\delta \cdot m_i(r_c) + (1-\delta) \cdot m_i(r_c))}{\gamma \cdot \mid r_c \mid + (1-\gamma) \cdot \mid r_f \mid} \tag{6}$$

The harmonic mean is calculated by the formula in equation 7.

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot P} \tag{7}$$

METEOR also accounts for word order differences and gaps by scaling F_{mean} by the fragmentation penalty (Pen). The fragmentation penalty (Pen) in equation 8 is computed by using the total number of matched words (m) and the number of chunks (ch).

$$Pen = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta} \tag{8}$$

The final score is then:

$$Score = (1 - Pen) \cdot F_{mean} \tag{9}$$

The parameters α, β, γ, δ and $w_i ... w_n$ are parameters that can be used for tuning METEOR for a given task.

3.2 Cosine Similarity

Cosine similarity is a similarity measure that can handle the fact that very similar documents (in our case sentences) may have different lengths. The cosine similarity of two documents is calculated by deriving a vector ($\vec{V}$) for each sentence or document d, denoted as $\vec{V}(d)$[2]. The set of documents in a collection is viewed as a set of vectors in a vector space, each term (meaning a word) having its own axis. By this kind of representation the initial ordering of terms in the document is lost, since cosine similarity does not incorporate context.

The cosine of two vectors can be derived by using the Euclidean dot product formula:

$$a \cdot b = \mid a \mid\mid b \mid \cos\theta \tag{10}$$

Derived from the formula in (10) the similarity between two documents d_1 and d_2 can be computed by the cosine similarity of their vector representations $\vec{V}(d_1)$ and $\vec{V}(d_2)$.

$$\cos(\theta) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{\mid \vec{V}(d_1) \mid\mid \vec{V}(d_2) \mid} \tag{11}$$

The numerator in (11) represents the dot product of the vectors $\vec{V}(d_1)$ and $\vec{V}(d_2)$ and is defined as shown in equation (12).

$$\vec{V}(d_1) \cdot \vec{V}(d_2) = \sum_{i=1}^{n} \vec{V_i}(d_1) \times \vec{V_i}(d_2) \tag{12}$$

The denominator corresponds to the product of the Euclidean length of the vectors $\vec{V_i}(d_1)$ and $\vec{V_i}(d_1)$.

$$\mid \vec{V}(d_1) \mid = \sqrt{\sum_{i=1}^{n} \vec{V_i}(d_1)} \tag{13}$$

The vectors are length normalised by the formulas in (13) and (14).

$$\mid \vec{V}(d_2) \mid = \sqrt{\sum_{i=1}^{n} \vec{V_i}(d_2)} \tag{14}$$

3.3 `ZWICKEL`: A Regression-based Metric

Similar to the Semantic Textual Similarity (STS) and MT Quality Estimation approaches (Scarton et al., 2015), we treat the MT metric task as a regression task with the aim of learning a Bayesian Ridge function that maps the cosine similarity feature to the target METEOR score.

A Bayesian Regressor finds a maximum a posteriori solution under a Gaussian prior *N* over the parameters w with the precision of λ^{-1}. The α and λ parameters are treated as random variables estimated from the data.

$$p(y|X, w, \alpha) == N(y|X, w, \alpha) \tag{15}$$

[2]The normalization of the terms in the vector is computed by using using $TF * IDF$

The Bayesian Ridge estimates a probabilistic regression model with a zero-mean prior for the parameter w, given by a spherical Gaussian:

$$p(w|\lambda) = N(w|0, \lambda^{-1} I_p) \quad (16)$$

Without the caveats of mathematical argot, we refer to the cosine similarities as *X*, and to the METEOR scores as *Y*. We aim to learn a regressor that outputs the paraphrase and synonym METEOR scores using the cosine similarities, without the paraphrase/synonym tables. Essentially, this leads to a language independent METEOR measure based on cosine similarity between translation and reference vectors.

3.4 COMET: A Combination of METEOR and ZWICKEL

We noticed that the outputs of the basic ZWICKEL score is conservative and does not allow an extreme 0.0 or 1.0 score unlike the METEOR score. Thus, we created a "*switch-like metric*", COMET, that treat the METEOR scores as oracle when METEOR reports 0.0 or 1.0 scores, otherwise it falls back to ZWICKEL.

4 Experiments

This year's USAAR submission to the WMT metrics shared task concentrated on evaluating translations into German and into English, assigning a score both at sentence and system level.

4.1 Training Data

For training our system we used the available data from the previous WMT shared tasks by conflating them into a single data set[3]. The into German set consisted of 359545 sentence pairs and the into English set consisted of 1194017 sentence pairs.

4.2 Test Data

The test data for our evaluation metrics consist of all system outputs from this year's translation task performed on the newstest2015 data set. Depending on the source language the data sets consist of a different number of sentences. Into English we evaluated MT systems having the following source languages:

- Czech with 10 system submissions and 2655 translated sentences per system
- German with 13 system submissions and 2168 translated sentences per system
- Finnish with 14 system submissions and 1369 translated sentences per system
- Russian with 13 system submissions and 2817 translated sentences per system

Into German we evaluated 16 systems with 2168 translated sentences per system.

Based on the sentence scores we provided also a system score for each language pair. The system score was calculated by using different means (median, arithmetic mean, arithmetic geometric mean, harmonic mean and root squared mean) for each proposed metric.

4.3 USAAR's Submission to the WMT2015 Metrics Shared Task

In order to evaluate the efficacy of our method we contributed with three systems to the metrics task:

- COSINE: the raw document embedding similarity, i.e. *sim(hyp, ref)*
- ZWICKEL: the cosine-based metric outputs from the regressor described above
- COMET: the combination of ZWICKEL outputs from the regressor and METEOR

5 Evaluation

All submissions to the metrics task were evaluated[4] at system level by computing their Pearson correlation coefficient with human judgements. For the evaluation of translations into English our best submission is COMET, achieving on average a correlation coefficient of 0.788 ± 0.026. For the evaluation of translations from English into German, COMET is again our best submission with a correlation coefficient of 0.448 ± 0.40.

Table 5 shows the system-level Pearson correlation coefficient for COSINE, ZWICKEL and COMET[5] for each language pair into English and for the language pair English-German.

Spearman's correlation coefficient was also computed, but just the average over all language

[3] We have compiled the WMT08-15 metrics task data sets into a single python-readable library that is easily accessible at https://github.com/alvations/warmth.

[4] The numbers reported in this section are provided by the organisers of the WMT2015 metrics shared task

[5] For the translations into English the system-level score is the root mean square of the sentence-level scores. For the translations from English into German the best system-level scores are achieved by the arithmetic geometric mean of the sentence-level scores.

	Pearson Correlation Coefficient		
Language pair	COSINE	ZWICKEL	COMET
Finnish-English	NaN	-0.093±0.043	0.834±0.023
German-English	0.008±0.052	0.286±0.052	0.847±0.027
Czech-English	0.912±0.013	0.406±0.031	0.896±0.014
Russian-English	NaN	0.264±0.052	0.603±0.041
English-German	NaN	-0.232±0.044	0.448±0.040

Table 1: Pearson correlation coefficient for COSINE, ZWICKEL and COMET.

	Spearman's Correlation Coefficient		
Average	COSINE	ZWICKEL	COMET
into English	0.122±0.079	0.066±0.087	0.665±0.069
into German	0.084±0.084	-0.235±0.069	0.588±0.072

Table 2: System-level Spearman's correlation coefficient for COSINE, ZWICKEL and COMET.

pairs into English and into German. From the results in Table 5 we notice that COMET was the metric performing best for both translations into English and German, achieving a coefficient of 0.665±0.069 for translations into English and 0.588±0.072 for translations from German into English.

6 Conclusion

This paper presents USAAR's submission to the WMT2015 metrics shared task. Our aim of our submission was a language independent method for predicting MT adequacy based on the semantic similarity between hypothesis and reference translation by using document embeddings. We contributed with three evaluation metrics, COMET, a combination of a cosine-based metric and METEOR, being the one correlating best with the human evaluators.

Previous studies have shown that METEOR systematically underestimate the quality of the translations (Vela et al., 2014b). Future work on our approach using document embeddings and cosine similarities could be used to also predict different scores (i.e. other than METEOR). Additionally, further experiments on document/word embeddings would be beneficial to find the best-fit solution for the cosine similarity calculation between a machine translation and its reference translation.

Acknowledgements

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 317471.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, August.

Piyush Arora, Chris Hokamp, Jennifer Foster, and Gareth Jones. 2015. DCU: Using Distributional Semantics and Domain Adaptation for the Semantic Textual Similarity SemEval-2015 Task 2. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 143–147, Denver, Colorado, June.

Alberto Barrón-Cedeño, Lluís Màrquez, Maria Fuentes, Horacio Rodriguez, and Jordi Turmo. 2013. UPC-CORE: What Can Machine Translation Evaluation Metrics and Wikipedia Do for Estimating Semantic Textual Similarity? In *2nd Joint Conference on Lexical and Computational Semantics (SEM)*, pages 143–147, Atlanta, Georgia, USA, June.

Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. 2011. Learning Structured Embeddings of Knowledge Bases. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2012. Joint learning of words and meaning representations for open-text semantic parsing. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 127–135, La Palma, Canary Islands, April.

Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. 2015. Learning Distributed Word Representations for Natural Logic Reasoning. In *Proceedings of the Association for the Advancement of Artificial Intelligence Spring Symposium (AAAI)*, pages 10–13, March.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the 2nd Workshop on Statistical Machine Translation (WMT)*, pages 136–158.

Danqi Chen and Christopher Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October.

Elisabet Comelles and Jordi Atserias. 2014. VERTa Participation in the WMT14 Metrics Task. In *Proceedings of the 9th Workshop on Statistical Machine Translation (WMT)*, pages 368–375, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd International Conference on Human Language Technologies (HLT)*, pages 138–145.

Christian Federmann. 2012. Appraise: An Open-Source Toolkit for Manual Evaluation of Machine Translation Output. *PBML*, 98:25–35, 9.

Meritxell Gonzàlez, Alberto Barrón-Cedeño, and Lluís Màrquez. 2014. IPA and STOUT: Leveraging Linguistic and Source-based Features for Machine Translation Evaluation. In *Proceedings of the 9th Workshop on Statistical Machine Translation (WMT)*, pages 394–401, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Ana Guerberof. 2009. Productivity and Quality in the Post-editing of Outputs from Translation Memories and Machine Translation. *International Journal of Localization*, 7(1).

Aaron L. F. Han, Derek F. Wong, and Lidia S. Chao. 2012. LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors. In *Proceedings of COLING 2012: Posters*, pages 441–450, Mumbai, India, December.

Pingping Huang and Baobao Chang. 2014. SSMT:A Machine Translation Evaluation View To Paragraph-to-Sentence Semantic Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 585–589, Dublin, Ireland, August.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 873–882.

Thomas K. Landauer and Susan T. Dutnais. 1997. A Solution to Platos Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *PSYCHOLOGICAL REVIEW*, 104(2):211–240.

Vladimir Iosifovich Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710.

Linguistics Data Consortium. 2005. Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations.

Chi-Kiu Lo and Dekai Wu. 2011a. MEANT: An Inexpensive, High-accuracy, Semi-automatic Metric for Evaluating Translation Utility Based on Semantic Roles. In *Proceedings of the 49th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 220–229.

Chi-Kiu Lo and Dekai Wu. 2011b. SMT vs. AI redux: How Semantic Frames Evaluate MT More Accurately. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*.

Chi-Kiu Lo and Dekai Wu. 2011c. Structured vs. Flat Semantic Role Representations for Machine Translation Evaluation. In *Proceedings of the 5th Workshop on Syntax and Structure in Statistical Translation (SSST)*.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. 2012. Fully Automatic Semantic MT Evaluation. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT)*, pages 243–252, Montréal, Canada, June. Association for Computational Linguistics.

Tucker Maney, Linda Sibert, Dennis Perzanowski, Kalyan Gupta, and Astrid Schmidt-Nielsen. 2012. Toward Determining the Comprehensibility of Machine Translations. In *Proceedings of the 1st PITR*, pages 1–7.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv*, 1301.3781.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, November.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1341–1351, Sofia, Bulgaria, August.

Miguel Rios, Wilker Aziz, and Lucia Specia. 2012. UOW: Semantically Informed Text Similarity. In *The 1st Joint Conference on Lexical and Computational Semantics (SEM)*, pages 673–678, Montréal, Canada, June.

Carolina Scarton, Marcos Zampieri, Mihaela Vela, Josef van Genabith, and Lucia Specia. 2015. Searching for Context: a Study on Document-Level Labels for Translation Quality Estimation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 121–128, Antalya, Turkey, May.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In *Proceedings of the 4th Workshop on Statistical Machine Translation (WMT)*, pages 259–268.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. DLS@CU: Sentence Similarity from Word Alignment and Semantic Vector Composition. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 148–153, Denver, Colorado, June.

Liling Tan and Francis Bond. 2013. Xling: Matching query sentences to a parallel corpus using topic models for wsd. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 167–170, Atlanta, Georgia, USA, June.

Liling Tan, Rohit Gupta, and Josef van Genabith. 2015a. Usaar-wlv: Hypernym generation with deep neural nets. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 932–937.

Liling Tan, Carolina Scarton, Lucia Specia, and Josef van Genabith. 2015b. USAAR-SHEFFIELD: Semantic Textual Similarity with Deep Regression and Machine Translation Evaluation Metrics. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 85–89, Denver, Colorado, June.

Liling Tan. 2013. Examining Crosslingual Word Sense Disambiguation. Master's thesis, Nanyang Technological University.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Alexander Zubiaga, and Hassan Sawaf. 1997. Accelerated DP Based Search For Statistical Translation. In *Proceedings of the EUROSPEECH*, pages 2667–2670.

Mihaela Vela and Josef van Genabith. 2015. Reassessing the WMT2013 Human Evaluation with Professional Translators Trainees. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 161–168, May.

Mihaela Vela, Anne-Kathrin Schumann, and Andrea Wurm. 2014a. Beyond Linguistic Equivalence. An Empirical Study of Translation Evaluation in a Translation Learner Corpus. In *Proceedings of the EACL Workshop on Humans and Computer-assisted Translation (HaCat)*, pages 47–56, April.

Mihaela Vela, Anne-Kathrin Schumann, and Andrea Wurm. 2014b. Human Translation Evaluation and its Coverage by Automatic Scores. In *Proceedings of the Language Resources and Evaluation Conference Workshop on Automatic and Manual Metrics for Operational Translation Evaluation (MTE)*, pages 20–30, May.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of Language Resources and Evaluation Conference*, pages 46–50, Valletta, Malta.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. TAKELAP: Systems for Measuring Semantic Text Similarity. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (SEM)*, pages 441–448.

Sandra Weiss and Lars Ahrenberg. 2012. Error Profiling for Evaluation of Machine-translated Text: a Polish-English Case Study. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC)*, pages 1764–1770.

Marcos Zampieri and Mihaela Vela. 2014. Quantifying the Influence of MT Output in the Translators Performance: A Case Study in Technical Translation. In *Proceedings of the EACL Workshop on Humans and Computer-assisted Translation (HaCat)*, pages 93–98, April.