Using Contextual Information for Machine Translation Evaluation

Marina Fomicheva, Núria Bel
IULA, Universitat Pompeu Fabra
Barcelona, Spain
marina.fomicheva@upf.edu, nuria.bel@upf.edu

Abstract

Automatic evaluation of Machine Translation (MT) is typically approached by measuring similarity between the candidate MT and a human reference translation. An important limitation of existing evaluation systems is that they are unable to distinguish candidate-reference differences that arise due to acceptable linguistic variation from the differences induced by MT errors. In this paper we present a new metric, UPF-Cobalt, that addresses this issue by taking into consideration the syntactic contexts of candidate and reference words. The metric applies a penalty when the words are similar but the contexts in which they occur are not equivalent. In this way, Machine Translations (MTs) that are different from the human translation but still essentially correct are distinguished from those that share high number of words with the reference but alter the meaning of the sentence due to translation errors. The results show that the method proposed is indeed beneficial for automatic MT evaluation. We report experiments based on two different evaluation tasks with various types of manual quality assessment. The metric significantly outperforms state-of-the-art evaluation systems in varying evaluation settings.

Keywords: Machine Translation, Evaluation, Local Context, Alignment

1. Introduction

Automatic evaluation of Machine Translation (MT) is based on the idea that the closer the MT output is to a human reference translation, the higher its quality. Thus, the task is typically approached by measuring some kind of similarity between the MT (also called candidate translation) and a reference translation. Most widely used evaluation systems follow a simple strategy of counting the number of matching words and word strings in the MT and a human reference. For example, the well known metric BLEU (Papineni et al., 2002) measures the number of word n-grams in the candidate translation that are also present in the reference. This approach, however, is not reliable since the same source sentence can be correctly translated in many different ways. The fact that the MT output does not match one of the possible translation options is not necessarily indicative of low MT quality.

Substantial work has focused on improving reference-based evaluation with various strategies: use of additional references (Albrecht and Hwa, 2008; Madnani and Dorr, 2013; Fomicheva et al., 2015a), integration of linguistic information (Padó et al., 2009; Giménez and Màrquez, 2010; Comelles et al., 2012; Denkowski and Lavie, 2014; Guzmán et al., 2014) and use of machine learning techniques (Gupta et al., 2015; Herrera et al., 2015). Despite important achievements, automatic evaluation is still a poor substitute for manual quality assessment. The correlation between the metrics' scores and human judgments of translation quality at sentence level continues to be low. The reason is that when comparing candidate and reference translations, the metrics are not able to distinguish acceptable linguistic variation from the differences that are indicative of MT errors. In this work we propose to use local context to discriminate between acceptable and non-acceptable differences. Thus, variation between the MT and a human translation can be considered meaning-preserving if they contain semantically similar words and the words occur in syntactically equivalent contexts. In case of translation errors either the lexical choice is inappropriate or the syntactic contexts of the matching words are not equivalent (word order errors, wrong choice of function words, etc.).

We have developed a new evaluation system, UPF-Cobalt, that exploits contextual information for estimating to what extent lexical matches between candidate and reference words are indicative of sentence-level translation quality. Following the success of Meteor (Denkowski and Lavie, 2014) we adopt a two-stage approach to evaluation. The MT output is first word-aligned to the reference and then scored based on the proportion of aligned words.

The novel contribution of our method is that a score for each pair of aligned words is calculated combining the information on their lexical similarity with the difference of their syntactic contexts, if any. The number and the syntactic functions of the context words are taken into consideration. In this way, the metric can make fine-grained distinctions regarding the relative importance of the differences between the MT and the reference translation. Furthermore, we increase the coverage of the cases of acceptable differences. At lexical level distributed representations of words (Mikolov et al., 2013) are used in order to identify contextual synonyms. At syntactic level, we take advantage of the classes of equivalent dependency types proposed by Sultan et al. (2014).

Using contextual information with the aforementioned enhancements helps to distinguish Machine Translations (MTs) that are different from the human translation and still essentially correct from those that share a high number of words with the reference but alter the meaning of the sentence due to translation errors.

We conduct experiments with the data from two different evaluation tasks with various types of human judgments of MT quality provided. The metric achieves competitive results in varying evaluation settings, including the well known Metrics Task at the Association for Computational

Linguistics (ACL) Workshop on Statistical Machine Translation (WMT) where it was ranked among the 4 best performing systems (Macháček and Bojar, 2015). Experimental results thus confirm that the integration of syntactic context into word-level candidate-reference comparison is indeed highly beneficial for MT evaluation.

The rest of this paper is organized as follows. Section 2 examines relevant pieces of related work. Section 3 describes our evaluation metric. In Section 4 we present the experiments and analyze the results. Finally, conclusions are given in Section 5.

2. Related Work

Evaluation systems based on surface-level similarity between the MT and a reference translation penalize acceptable differences induced by the use of semantically equivalent expressions that do not match in their surface forms. At the same time, the matches between the words that happen to have the same form but play totally different roles in the corresponding sentences incorrectly increase the evaluation score.

The issue of acceptable variation has been addressed by using additional references. It has been shown that the performance of BLEU is improved when various human translations are used as benchmarks (Dreyer and Marcu, 2012). Having multiple human references is expensive. Albrecht and Hwa (2008) use pseudo-references as additional source of information. Data-driven (Owczarzak et al., 2006) and rule-based paraphrase generation (Fomicheva et al., 2015a) have also been explored. These approaches, however, fail to estimate the varying impact of different types of candidate-reference mismatches on MT quality.

An alternative strategy is to refine the comparison between the candidate MT and the available human translation. Meteor (Denkowski and Lavie, 2014) allows for stem, synonym and paraphrase matches, thus addressing the problem of acceptable variation at lexical level. Liu and Gildea (2005) propose a series of syntactic features based on the degree of overlap between the syntactic trees of candidate and reference translations.

Translation quality is a complex object involving different aspects. A number of successful approaches, therefore, combine different types information. Thus, Giménez and Màrquez (2010) propose a combination of specialized similarity measures operating at various linguistic levels (lexical, syntactic and semantic). Guzmán et al. (2014) further enrich this metric set with discourse level information, obtaining a marginal improvement. Our work follows this line of research. But instead of adding new sources of linguistic evidence, we propose a refined way of combining lexical and syntactic similarity at word level, that allows to estimate the impact of candidate-reference differences on sentence-level quality.

3. UPF-Cobalt

For a meaningful comparison, not only the number but also the nature of the correspondences between the words in the MT and the human reference must be taken into consideration. Therefore, we have chosen to perform the evaluation in two stages. First, the MT output is aligned to the reference. Next, the MT is scored taking into account both the number of aligned words and their roles in the corresponding sentences.

3.1. Alignment

In our setting, it is important to establish the relations between candidate and reference words correctly. Research in the area of monolingual alignment demonstrates that exploiting syntactic context to discriminate between possible alignments results in significant improvements (Thadani et al., 2012). The alignment module of UPF-Cobalt builds on an existing system Monolingual Word Aligner (MWA) which takes context information into account and has been shown to significantly outperform state-of-the-art results (Sultan et al., 2014).

3.1.1. Monolingual Word Aligner

MWA makes alignment decisions based on lexical similarity and contextual evidence. The lexical similarity component identifies the word pairs that are possible candidates for alignment. Two levels of similarity are defined. In addition to the exact or lemma match, Paraphrase Database (Ganitkevitch et al., 2013) of lexical and phrasal paraphrases is employed to recognize semantically similar words.

Context words are considered as evidence for alignment if they are lexically similar and have the same or equivalent syntactic relations with the words to be aligned. Syntactic equivalence is established through a mapping between different syntactic functions that instantiate the same semantic relation. Some examples of such functions are: possession modifier and noun compound modifier, indirect object and prepositional modifier, relative clause modifier and reduced non-finite verbal modifier, nominal subject of an active clause and by-agent in a passive clause. See Sultan et al. (2014) for a complete list of functions. We use this mapping at the scoring stage in order to avoid penalizing syntactic variation.

3.1.2. Distributional Similarity

To get better lexical coverage, we integrate two additional levels to the MWA's lexical similarity component. In addition to the Paraphrase Database, UPF-Cobalt employs WordNet synonyms (Miller and Fellbaum, 2007) and distributed word representations (Mikolov et al., 2013). WordNet and paraphrase databases are commonly used in MT evaluation for dealing with lexical variation. By contrast, to the best of our knowledge, distributional similarity has not yet been exploited.

Distributional semantic models (Baroni and Lenci, 2010) have been shown to perform well across a variety of lexical similarity tasks. They are grounded on distributional hypothesis (Harris, 1954) that states that semantic similarity between two words can be modeled as a function of the degree of overlap between their contexts. In this framework, words are represented as vectors in which each entry is a measure of association between the word and a particular context. The similarity between two given words is then computed using some distance measure on the corresponding vectors.

Using distributional similarity in combination with contextual information is highly beneficial for MT evaluation, since it helps to identify quasi-synonyms, i.e. words that can be considered synonymous only given the similarity of their contexts. Consider the following example.

Ref: I understand that the Council has also signalled its agreement in principle.
MT: I understand that the Council has also given its consent in principle.

The correspondence between the words "agreement" and "consent" can be easily established with the help of common lexical similarity resources such as WordNet. This is not the case, however, with the words "signalled" and "given", which can be considered semantically equivalent only given the equivalence of their contexts.
Recently it has been proposed to represent words as dense vectors derived by various training methods inspired from neural-network language modeling (Mikolov et al., 2013). These representations, referred to as word embeddings, have been shown to outperform previous approaches (Baroni et al., 2014). We use dependency-based word embeddings developed by Levy and Goldberg (2014) and cosine similarity as a distance measure. The words that have cosine similarity higher than a threshold[1] and at least one pair of exact matching content words in their contexts are considered candidates for alignment.

3.2. Scoring

At the scoring stage we want to know if the word correspondences identified by the aligner are actually indicative of MT quality. UPF-Cobalt calculates a score for each pair of aligned words as a combination of their lexical similarity and a context penalty which measures the difference in their syntactic contexts.

3.2.1. Lexical Similarity

The values for different types of lexical similarity are defined as follows: same word forms - 1.0, lemmatizing or stemming - 0.9, WordNet synsets - 0.8, paraphrase database - 0.6, distributional similarity - 0.5. These values were established heuristically, depending on the accuracy of the lexical resource that was used for aligning the corresponding words.

3.2.2. Context Penalty

Context penalty is applied at word level to identify cases where the words are aligned (i.e. lexically similar) but play different roles in the sentences and therefore should contribute less to the sentence-level evaluation score. Thus, for each pair of aligned words, the words that constitute their syntactic contexts are compared. The syntactic context of a word is defined as its head and dependent nodes in a dependency graph.[2] Both the context words and their dependency labels are compared.

The following issues are taken into consideration when measuring contextual differences. First, mistranslating the words with argument functions (subject, direct object, prepositional object, etc.) changes the context to a greater extent than dropping a determiner or an adjunct. Therefore, context words are assigned different weights depending on the relative importance of their syntactic functions. Second, to account for the possible equivalence of certain syntactic relations we use the mapping described in Section 3.1.1. As shown by Fomicheva et al. (2015a), syntactic variation is a regular source of differences between human reference and MT. By taking it into consideration, we avoid penalizing perfectly acceptable MTs that contain different syntactic structures but are semantically similar to the reference translation. Finally, the number of context words is taken into account assuming that a candidate-reference difference involving a word with more syntactic dependents has a higher impact on the MT quality.
For each pair of aligned words, t in the candidate translation and r in the reference translation, the context penalty is calculated as follows:

$$\begin{aligned} CP(t,r) &= \frac{\sum_{1..i} w(C_i^*)}{\sum_{1..i} w(C_i)} \times ln\left(\sum_{1..i} w(C_i) + 1\right) \\ Pen(t,r) &= \frac{2}{1+e^{-CP(t,r)}} - 1 \end{aligned} \tag{1}$$

Where CP stands for context penalty, C refers to the words that belong to the syntactic context of the word r and C_i^* refers to the context words that are **not** equivalent.[3] For the words to be equivalent two conditions are required to be met: a) they must be aligned and b) they must be found in the same or equivalent syntactic relation with the word r.
The weights w that reflect the relative importance of the dependency functions of the context words are defined as follows: argument/complement functions - 1.0, modifier functions - 0.8, specifier functions - 0.2.
The number of context words is taken into consideration assuming that the higher the number of syntactic dependents a word has, the higher will be the impact of a candidate-reference difference involving this word. We use the natural logarithm of the weighted count of context words, since this impact saturates above a threshold. Thus, a context difference receives a higher value when the number of context words is high (it is not the same translating zero words out of one and zero words out of ten), while limiting the increase if the number of context words continues to grow (the difference between translating six words out of eight and eight words out of ten is less relevant). To obtain the final value for context penalty (Pen), CP is normalized from 0 to 1 using logarithmic function. Then, given the information on lexical similarity and contextual differences, the score for each pair of aligned words is:

$$score(t,r) = LexSim(t,r) - Pen(t,r) \tag{2}$$

Finally, sentence-level score is calculated as a weighted

[1] Based on data observation, we currently define the threshold as 0.25.

[2] Stanford dependency parser (de Marneffe et al., 2006) is used to extract the dependencies.

[3] Context penalty is calculated both on reference and on candidate sides and the resulting values are averaged.

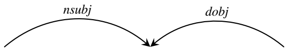
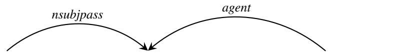
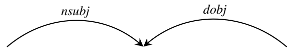
 Ref: The government has discussed the document .	Equivalent dep. types	Scores	
		UPF-Cobalt	Meteor
 MT1: The document was discussed by the government .	*nsubj* ≈ *agent* *dobj* ≈ *nsubjpass*	0.908	0.408
 MT2: The document has discussed the government .	*nsubj* ≠ *dobj* *dobj* ≠ *nsubj*	0.642	0.464

Table 1: Example of candidate and reference translations with the corresponding Meteor and UPF-Cobalt scores

combination of precision and recall over the sum of the individual scores for aligned candidate and reference words. We note that word-level context penalty captures the propagation of translation errors. If the mistranslated word have many syntactic dependents all of them will receive a context penalty, which will strongly affect the score at sentence level. By contrast, if the error involves a word that has few syntactic dependents its impact will be low.

To appreciate the advantages of the method proposed, Table 1 provides a qualitative comparison of the performance of UPF-Cobalt and Meteor. Here MT1 is assigned a low score by Meteor due to the change in surface word order. UPF-Cobalt correctly assigns a high score to this sentence. All the content words are aligned and no context penalty is applied as the syntactic contexts of the aligned words are equivalent. Thus, *agent* relation in the candidate translation is equivalent to nominal subject relation (*nsubj*) in the reference, and subject of a passive clause (*nsubjpass*) in the candidate corresponds to the direct object (*dobj*) in the reference.

By contrast, Meteor assigns a higher score to MT2 because of a matching auxiliary verb which in this case is not indicative of candidate-reference semantic similarity. MT2 is assigned a much lower score by UPF-Cobalt. Although all content words are matched they occur in different contexts and receive a high context penalty (0.90 for the main verb ”discussed” and 0.80 for the arguments ”government” and ”documents”). Thus, UPF-Cobalt is capable of distinguishing the use of equivalent constructions (active/passive alternation) from translation errors. The context penalty values calculated for each pair of aligned words can be used for locating translation errors.

Examples of acceptable syntactic variation are frequently found in professional human translation (Ahrenberg, 2005). Translators often introduce optional changes to the original sentence in order to adhere to specific principles of target language use, resulting in the existence of various possible translations with a varying distance from the source sentence. If the available human reference contains optional changes with respect to the source, surface-level comparison is not informative, as the absence of such changes is not indicative of low MT quality.

4. Experiments

The performance of evaluation systems is typically assessed by comparing the scores produced by the metrics with the results of manual MT evaluation. Over the years, various settings have been developed for human evaluation in order to increase its reliability. Traditionally, MT is evaluated in terms of absolute quality, on a multi-point scale. The two main criteria used for absolute scoring are adequacy and fluency (Linguistic Data Consortium, 2005). Adequacy measures how much of the meaning of the source sentence (or human reference translation) is preserved in the MT. Fluency refers to the well-formedness of the translation. This type of evaluation is thus based on the defining properties of the translation and constitutes a powerful and intuitive instrument for assessing MT quality. Measuring absolute quality on a interval level scale, however, presents a problem of low inter-annotator agreement. The scale is arbitrary and no precise instructions are given to the annotators. As a result, different judges may assign different scores for the same sentence.

To overcome this issue an alternative setting has been introduced, in which the judges are asked to rank different MTs of the same source sentences in terms of their relative quality (Callison-Burch et al., 2007). While this formulation of the task results in a higher inter-annotator agreement, it is less informative than absolute quality judgments.

It has been shown that the performance of automatic evaluation systems varies significantly depending on the type of human judgments and the error metric (Denkowski and Lavie, 2010). Different types of human judgments pose different challenges to automatic evaluation systems. Ranking can be more difficult when very similar MTs have to be compared, in which case fine-grained distinctions between different kind of errors have to be made. On the other hand, in the ranking task the scores produced by a metric are not assessed directly. Ranking judgments provide little insight regarding how well the magnitude of the differences in quality between the MTs of different source sentences is reflected in automatic evaluation.

MT can be evaluated at system or at sentence level. System-level evaluation is typically conducted by averaging sentence-level scores. It is useful for comparing the performance of different MT systems and allows identifying the advantages and limitations of MT strategies. Sentence-level evaluation is crucial for parameter tuning of statistical MT systems and provides fine-grained judgments of translation quality. Here we focus on sentence-level evaluation, since automatic evaluation at system level is largely considered a solved problem.

Metric	fr-en	fi-en	de-en	cs-en	ru-en	Avg τ
UPF-Cobalt	0.386	0.437	0.427	0.457	0.402	.422±.011
DPMFcomb(Yu et al., 2015)	0.395	0.445	0.482	0.495	0.418	.447±.011
BEER_Treepel(Stanojevic and Sima'an, 2015)	0.389	0.438	0.447	0.471	0.403	.429±.011
RATATOUILLE(Marie and Apidianaki, 2015)	0.398	0.421	0.441	0.472	0.393	.425±.010
BLEU(Papineni et al., 2002)	0.358	0.308	0.360	0.391	0.329	.349±.011
Meteor(Denkowski and Lavie, 2014)	0.380	0.406	0.422	0.439	0.386	.407±.012
Asiya(Giménez and Màrquez, 2010)	0.360	0.351	0.391	0.424	0.358	.377±.011

Table 2: Sentence-level evaluation results for WMT15 dataset in terms of Kendall rank correlation coefficient (τ)

We conduct experiments with different types of human judgments and show the robustness of our method in varying evaluation settings. See Fomicheva et al. (2015b) for a detailed analysis of the importance of different components of the metric.

4.1. Relative Quality

In this scenario human annotators are asked to judge translations in terms of their relative quality. We use the data from 2015 Workshop on Statistical Machine Translation (WMT). The dataset consists of source texts, human reference translations and the outputs from the participating MT systems, for five different language pairs. Manual evaluation was performed using an ordinal level scale. Annotators were presented with the source sentence, its human translation and the output of five MT systems and asked to rank the MTs from best to worst. Kendall rank correlation coefficient (τ) is used to measure the correlation between metrics' scores and human ranking. Specifically, we use the definition of Kendall τ presented in Macháček and Bojar (2015) which was the official measure for the WMT15 Metrics Task. Table 2 shows the results for all into-English translation directions.[4]

Our metric participated in the WMT15 Metrics Task and was ranked among the 4 best performing systems for sentence-level evaluation. Similar results were obtained for previous WMT workshops and are reported in Fomicheva et al. (2015b). For the sake of comparison the first group of results in Table 2 reproduces the correlations of the metrics that outperformed UPF-Cobalt at WMT15 Metrics Task. DPMFComb and RATATOUILLE use a learnt combination of the scores from different evaluation metrics, while BEER_Treepel employs leaning-to-rank approach to combine string-level and syntax-level features.

The second group of results corresponds to the baseline n-gram based evaluation system BLEU and a strong baseline Meteor that uses synonyms and paraphrases to address lexical variation. Also, we calculate the correlation for ULC system developed by Giménez and Màrquez (2010). This is a uniform linear combination of metrics based on various levels of linguistic information. At syntactic level, ULC uses the degree of overlap between dependency trees of candidate and reference translations, and is thus comparable to our approach.

First, we observe that UPF-Cobalt significantly outperforms the baseline systems, as well as the linguistically informed ULC metric, which considers lexical and syntactic aspects separately. As shown in Fomicheva et al. (2015b), the gain in performance is mainly due to the use of context penalty. Secondly, we note that the performance of the metric varies depending on the source language. The improvement over baseline systems is small for French-English and German-English. Our intuition is that the metric achieves better results when evaluating translations involving distant language pairs. In case of typologically related languages, the syntactic parser may assign acceptable structures to ill-formed MT outputs, thus increasing the noise when considering the equivalence of different syntactic functions.

4.2. Absolute Quality

To test the metric's performance on absolute quality judgments, we conduct experiments with the MTC-P4 Chinese-English dataset, produced by Linguistic Data Consortium (LDC2006T04). This dataset contains 919 source sentences from news domain, 4 reference translations and MTs generated by 10 translation systems. The translations produced by 6 of the systems were assigned quality scores following the Linguistic Data Consortium evaluation guidelines (Linguistic Data Consortium, 2005), based on fluency and adequacy criteria, on a 5-point scale. In total, human assessment is provided for 5,514 MT sentences.

Fluency and adequacy scores are normally averaged to obtain global quality scores. We report sentence-level Pearson correlation with the averaged scores, as well as for fluency and adequacy scores separately. We compare the performance of our metric with BLEU (Papineni et al., 2002) and Meteor (Denkowski and Lavie, 2014).

MTC-P4 dataset contains 4 different human reference translations. The metrics are evaluated in both single-reference and multi-reference scenarios. For the case when only one human reference is used, the reference is chosen at random and is the same for all the evaluation systems. BLEU was specifically designed to be used with multiple references. It counts the n-gram matches between the MT and any of the available human translations. To adapt Meteor and UPF-Cobalt to the multi-reference scenario, we follow a simple approach of selecting for each sentence the highest of the 4 sentence-level scores obtained with different references. (See Qin and Specia (2015) for a description of alternative strategies). The results are summarized in Table 3.

First, we observe that UPF-Cobalt outperforms BLEU and Meteor for adequacy, fluency and averaged human judgments, in single-reference as well as in multi-reference scenario. The differences between UPF-Cobalt and BLEU were found to be significant in all cases. The differences

[4] The 95% confidence intervals are obtained using bootstrap resampling method as reported in Macháček and Bojar (2015).

	Single Reference			Multiple References		
Metric	A	F	Avg	A	F	Avg
UPF-Cobalt	0.460	0.279	0.418	0.491	0.306	0.450
Meteor	0.450	0.262	0.405	0.488	0.302	0.447
BLEU	0.295	0.200	0.278	0.342	0.252	0.332

Table 3: Sentence-level evaluation results on MTC4-P4 dataset in terms of Pearson correlation with Adequacy (A), Fluency (F) and Averaged (Avg) adequacy and fluency judgments

between UPF-Cobalt and Meteor were found to be significant for fluency scores and average scores in the single-reference scenario.[5]

Secondly, all the metrics present a lower correlation for fluency. The reason is that neither of the reference-based evaluation systems explicitly addresses this aspect of translation quality. However, BLEU and Meteor are outperformed by UPF-Cobalt in terms of the correlation with fluency judgments. The reason is that syntactic similarity between MT and the reference reflects, although indirectly, the MT fluency. In general, adequacy and fluency are related aspects. If the MT is very similar to a reference, it is probably well-formed. Thus, a metric that is better for predicting adequacy will also show an improvement in predicting fluency judgments.

Finally, the results show that the benefit of using multiple references is much higher in the case of BLEU. This is not surprising, since the evaluation systems that allow for fuzzy matches between words and constructions are designed precisely to overcome the limitations of using single reference as benchmark. Furthermore, the difference between UPF-Cobalt and Meteor is minimal in the case of multi-reference evaluation. This suggests that the gain in performance achieved by UPF-Cobalt in the single-reference scenario is related to addressing the issue of acceptable variation between the candidate translation and the human reference.

5. Conclusion

We have presented an alignment-based MT evaluation metric, UPF-Cobalt, that combines the information on lexical similarity and the syntactic context of the words. We have shown that comparing the syntactic contexts of the aligned words helps to distinguish cases of acceptable linguistic variations from the differences that are indicative of MT errors. Our word-level context penalty allows for a better estimation of the impact of candidate-reference differences on the sentence-level MT quality. Also, we have enhanced existing methods for addressing meaning-preserving variation between candidate and reference translations by exploiting distributed word representations at lexical level and classes of equivalent dependency types at syntactic level.

We have performed experiments using two main types of human evaluation: absolute quality scores based on adequacy and fluency criteria and ranking of different MTs in terms of their relative quality. The results show that UPF-Cobalt achieves stable and highly competitive results in varying evaluation settings. The metric and the code are freely available for download at *https://github.com/amalinovskiy/upf-cobalt*.

6. Acknowledgments

This work was supported by IULA (UPF) and the FI-DGR grant program of the Generalitat de Catalunya.

7. Bibliographical References

Ahrenberg, L. (2005). Codified Close Translation as a Standard for MT. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*.

Albrecht, J. S. and Hwa, R. (2008). Regression for Machine Translation Evaluation at the Sentence Level. *Machine Translation*, 22(1-2):1–27.

Baroni, M. and Lenci, A. (2010). Distributional Memory: A General Framework for Corpus-based Semantics. *Computational Linguistics*, 36(4):673–721.

Baroni, M., Georgiana, D., and Kruszewski, G. (2014). Don't Count, Predict! A Systematic Comparison of Context-counting vs. Context-predicting Semantic Vectors. In *ACL (1)*, pages 238–247.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158. ACL.

Comelles, E., Atserias, J., Arranz, V., and Castellón, I. (2012). VERTa: Linguistic Features in MT Evaluation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 3944–3950.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *In LREC*, pages 449–454.

Denkowski, M. and Lavie, A. (2010). Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks. In *Proceedings of the Ninth Biennal Conference of the Association for Machine Translation in the Americas*.

Denkowski, M. and Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Dreyer, M. and Marcu, D. (2012). HyTER: Meaning-equivalent Semantics for Translation Evaluation. In *Proceedings of 2012 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 162–171. ACL.

[5] The Hotelling-Williams (Williams, 1959) test for dependent correlations was used for significance testing.

Fomicheva, M., Bel, N., and da Cunha, I. (2015a). Neutralizing the Effect of Translation Shifts on Automatic Machine Translation Evaluation. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015*, pages 596–607. Springer.

Fomicheva, M., Bel, N., da Cunha, I., and Malinovskiy, A. (2015b). UPF-Cobalt Submission to WMT15 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 373–379.

Ganitkevitch, J., Durme, B. V., and Callison-Burch, C. (2013). PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of the North American Chapter of the ACL*, pages 758–764.

Giménez, J. and Màrquez, L. (2010). Linguistic Measures for Automatic Machine Translation Evaluation. *Machine Translation*, 24(3-4):209–240.

Gupta, R., Orasan, C., and van Genabith, J. (2015). ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072.

Guzmán, F., Joty, S., Màrquez, L., and Nakov, P. (2014). Using Discourse Structure Improves Machine Translation Evaluation. In *ACL (1)*, pages 687–698.

Harris, Z. S. (1954). Distributional Structure. *Word*, 10(2-3):146–162.

Herrera, F. G., Joty, S. R., Marques, L., and Nakov, P. (2015). Pairwise neural machine translation evaluation. In *ACL(1)*, pages 805–814.

Levy, O. and Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *Proceedings of ACL (2)*, pages 302–308.

Linguistic Data Consortium. (2005). Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations. Technical report.

Liu, D. and Gildea, D. (2005). Syntactic Features for Evaluation of Machine Translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32.

Macháček, M. and Bojar, O. (2015). Results of the WMT15 Metrics Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273.

Madnani, N. and Dorr, B. J. (2013). Generating Targeted Paraphrases for Improved Translation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):40.

Marie, B. and Apidianaki, M. (2015). Alignment-based Sense Selection in METEOR and the RATATOUILLE Recipe. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 385–391.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Miller, G. and Fellbaum, C. (2007). Wordnet.

Owczarzak, K., Groves, D., Genabith, J. V., and Way, A. (2006). Contextual Bitext-derived Paraphrases in Automatic MT Evaluation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 86–93. ACL.

Padó, S., Galley, M., Jurafsky, D., and Manning, C. (2009). Robust Machine Translation Evaluation with Entailment Features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 1, pages 297–305. ACL.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318. ACL.

Qin, Y. and Specia, L. (2015). Insight into Multiple References in an MT Evaluation Metric. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 131–140. Springer.

Stanojevic, M. and Sima'an, K. (2015). BEER 1.1: ILLC UvA Submission to Metrics and Tuning Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 396–401.

Sultan, M. A., Bethard, S., and Sumner, T. (2014). Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence. *Transactions of the ACL*, 2:219–230.

Thadani, K., Martin, S., and White, M. (2012). A Joint Phrasal and Dependency Model for Paraphrase Alignment. In *Proceedings of COLING 2012: Posters*.

Williams, E. J. (1959). *Regression Analysis*, volume 14. Wiley, New York, USA.

Yu, H., Ma, Q., Wu, X., and Liu, Q. (2015). CASICT-DCU Participation in WMT2015 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 417–421.