UPF-Cobalt Submission to WMT15 Metrics Task

Marina Fomicheva **Núria Bel** **Iria da Cunha**
IULA, Universitat Pompeu Fabra
{marina.fomicheva,nuria.bel,iria.dacunha}@upf.edu

Anton Malinovskiy
Nuroa Internet S. L.
amalinovskiy@gmail.com

Abstract

An important limitation of automatic evaluation metrics is that, when comparing Machine Translation (MT) to a human reference, they are often unable to discriminate between acceptable variation and the differences that are indicative of MT errors. In this paper we present UPF-Cobalt evaluation system that addresses this issue by penalizing the differences in the syntactic contexts of aligned candidate and reference words. We evaluate our metric using the data from WMT workshops of the recent years and show that it performs competitively both at segment and at system levels.

1 Introduction

Current automatic MT evaluation methods are grounded on the following key idea: the closer an MT is to a professional Human Translation (HT), the higher its quality. Thus, metrics typically calculate evaluation scores based on some sort of similarity between machine and human translations. The performance of evaluation systems is in its turn evaluated by calculating the correlation with human judgments. Manual quality assessment can be conducted in various ways: adequacy and fluency scoring, calculating post-editing cost or post-editing time, error analysis, ranking, etc. In the latter case, humans are asked to compare the outputs of different MT systems and rank them in terms of quality. Ranking-based evaluation has gained a lot of attention in the recent years and is used in important evaluation campaigns such as the Metrics task at the Workshop on Machine Translation (WMT). This setting is preferred, since it has been shown to yield higher inter-annotator agreement than absolute quality assessment (Callison-Burch et al., 2007).

In our opinion, one of the main reasons why the correlation between automatic evaluation and human rankings is still not satisfactory is that metrics' scores are not discriminative enough to approximate human comparisons. Given various candidate translations of the same source sentence, all of them different from the reference, evaluation systems are often unable to determine which translation is better as they cannot tell apart candidate-reference differences related to acceptable linguistic variation and the differences induced by MT errors. Furthermore, if all candidate translations contain a number of translation errors, metrics fail to predict the human ranking because they make no estimation of the relative importance of different types of MT errors for the overall translation quality.

We suggest that the aforementioned limitations can be addressed by means of enhancing word comparison with contextual information. Variation between two translation options is acceptable if semantically similar words in the corresponding sentences occur in equivalent contexts. In case of translation errors either the lexical choice is inappropriate or the syntactic contexts of the words are different (incorrect choice of function words, word order errors, etc.).

Our evaluation metric, UPF-Cobalt[1] exploits contextual information by means of weighting the contribution of each pair of lexically similar words in candidate and reference translations depending on whether they occur in similar syntactic environments. Syntactic functions of the words in context are taken into consideration. In this way, more

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

[1] The metric is freely available for download at https://github.com/amalinovskiy/upf-cobalt.

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 373–379,
Lisboa, Portugal, 17-18 September 2015. ©2015 Association for Computational Linguistics.

fine-grained distinctions can be made regarding the relative importance of mistranslated material.

In this paper we present UPF-Cobalt submission to the WMT15 Metrics task. Experiments show that UPF-Cobalt achieves competitive results, both at segment and at system levels. On WMT14 data, our metric would have been ranked as second-best performing metric at segment level, and tied with the first best-performing metric at system level.

The rest of this paper is organized as follows. Section 2 describes UPF-Cobalt. In Section 3 we present the experiments and analyze the results. Section 4 examines relevant pieces of related work. Finally, in Section 5 we give the conclusions and suggest directions for future work.

2 Metric Description

Following MacCartney et al. (2006), we argue that for measuring sentence similarity and related tasks, identifying similar words and deciding on the relation between the two sentences should be kept separate. This is especially relevant for MT evaluation where system output may share a high number of similar words with the reference and still be grammatically ill-formed and totally unacceptable. Thus, not only the number but also the characteristics of the correspondences between candidate and reference words must be taken into consideration. Therefore, we follow a two-stage approach to evaluation. First, MT is aligned to the reference. Next, the candidate translation is scored taking into account both the number of aligned words and their roles in the corresponding sentences.

2.1 Monolingual Word Aligner

We assume that using better candidate-reference alignment results in better MT evaluation. Research in the area of monolingual alignment demonstrates that exploiting syntactic context to discriminate between candidate pairs for alignment significantly improves the results (MacCartney et al., 2008; Thadani et al., 2012; Yao et al, 2013; Sultan et al., 2014). The alignment module of UPF-Cobalt builds on an existing system Monolingual Word Aligner (MWA)[2] which takes context information into account and has been shown to significantly improve on state-of-the-art results (Sultan et al., 2014).

MWA exploits lexical similarity and contextual evidence to make alignment decisions. Lexical similarity component identifies possible candidates for alignment. In addition to exact and lemma match, Paraphrase Database (Ganitkevitch et al., 2013) of lexical and phrasal paraphrases is employed to recognize semantically similar words.[3]

We enhance MWA with additional lexical similarity resources to maximize the coverage of the alignment. In addition to the paraphrase database, UPF-Cobalt employs WordNet synsets (Miller and Fellbaum, 2007) and distributional similarity (Turney and Pantel, 2010). WordNet is commonly used in MT evaluation and related fields for dealing with lexical variation. By contrast, to the best of our knowledge, distributional similarity has not yet been exploited for the evaluation task.

We use publically available distributional similarity resource (Levy and Goldberg, 2014), which contains dependency-based word embeddings. To minimize the noise, we establish the following restrictions. To be considered candidates for alignment the words must have the cosine similarity higher than a threshold (based on data observation, we currently define it as 0.25). Also, they must have at least one pair of exact matching content words in their contexts.

Contextual evidence is used to choose the best alignment candidates and is defined as the number of similar words in the contexts of the words to be aligned. At syntactic level, the context is constituted by the head and dependent nodes in a dependency graph.[4] 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.

Sultan et al. (2014) have developed a list of mappings between different syntactic functions that instantiate the same semantic relation. Thus, for example, the dependency relation between subject and predicate in an active clause and by-agent and predicate in a passive clause are defined to be equivalent. We consider that this functionality is helpful for addressing syntactic variation in reference-based MT evaluation and reuse it for

[2] `https://github.com/ma-sultan/monolingual-word-aligner`.

[3] MWA does not support phrase-level alignments, but the framework is flexible enough to integrate them in the future.

[4] The dependencies are extracted with Stanford dependency parser (de Marneffe et al., 2006).

scoring.

2.2 Scoring Method

Given a candidate-reference alignment, we further need to know if the correspondences identified at the alignment stage are actually indicative of MT quality. UPF-Cobalt computes a score for each pair of aligned words as a combination of their lexical similarity and the differences of the syntactic contexts in which the words occur.

Lexical Similarity. The weights for different types of lexical similarity are established heuristically, depending on the accuracy of the lexical resource that was used for aligning them:[5]

- Word form: 1.0
- Lemma or stem: 0.9
- WordNet synsets: 0.8
- Paraphrase database: 0.6
- Distributional similarity: 0.5

Context Penalty. Context penalty is applied in cases where aligned words play different roles in the corresponding sentences. For each pair of aligned nodes (*h*) in the candidate translation and (*r*) in the reference translation context penalty is calculated as follows:

$$
\begin{aligned}
CP(h,r) &= \frac{\sum_{1..i} w(c_i)}{count(c)} \times ln(count(c)+1) \\
w(c_i) &= \begin{cases} 0, & \text{if } c_i \in |A| \\ w(dep(c_i)), & \text{otherwise} \end{cases}
\end{aligned}
\tag{1}
$$

Where (*c*) refers to the words that belong to the syntactic context of the reference word (*r*) (immediate neighbors in the dependency graph).[6] If the context word is found in the set of aligned word pairs $|A|$ and its counterpart in the candidate translation has the same or equivalent syntactic relation with the word (*h*), the weight $w(c_i)$ equals to 0. Otherwise, the weight is defined according to the relative importance of the dependency function of the context word. Intuitively, mistranslating or omitting words with syntactic functions that correspond to arguments alters the context to a greater extent than dropping a determiner or an adjunct. We define three groups of syntactic functions accordingly and establish the corresponding weights as follows:

- Arguments and complements: 1.0
- Modifiers and adjuncts: 0.8
- Specifiers and auxiliaries: 0.2

The natural logarithm of $count(c)$ in Formula (1) gives a higher value to the contextual difference when the number of context words is high, while limiting the increase if the number of context words continues to grow. The final value of context penalty is normalized from 0 to 1 using logarithmic function:

$$Pen(h,r) = 2 \times \frac{1}{1+e^{-CP(h,r)}} \tag{2}$$

Given the values of lexical similarity and context penalty, the score for each pair of aligned word is defined as follows:

$$a(h,r) = LexSim(h,r) - Pen(h,r) \tag{3}$$

Sentence-level score is then calculated as a weighted combination of precision and recall over the sum of the scores for aligned candidate and reference words. To obtain system-level scores, we computed the ratio of sentences in which each system was assigned the highest sentence-level score by our metric.

3 Experiments

We conduct experiments with the data from WMT13 and WMT14 Metrics tasks (Macháček and Bojar, 2013; Macháček and Bojar, 2014). To evaluate our metric's performance at segment level, we use Kendall's Tau correlation (τ) with human rankings, as defined in (Macháček and Bojar, 2014). At system level, we use Pearson correlation coefficient (r). Table 1 presents the results averaged over all into-English translation directions. For the sake of comparison, we provide the results for the best performing metrics that participated in WMT13 and WMT14 Metrics tasks, as well as baseline metrics BLEU (Papineni et al., 2002) and Meteor (Denkowski and Lavie, 2014).

As shown in Table 1, our approach is competitive (UPF-Cobalt would have been ranked as the best performing metric on WMT13 data and as the second best on WMT14 data) and generalizes well

[5] We experimented with optimizing the weights for different types of lexical similarity, as well as for the classes of dependency functions discussed below. However, the optimization gave approximately the same values, showing that our intuition was essentially correct.

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

	Segment-level		System-level	
Metric	WMT13	WMT14	WMT13	WMT14
DiscoTK-Party-Tuned (Guzman et al., 2014)	-	0.386	-	0.944
BEER (Stanojević and Sima'an, 2014)	-	0.362	-	-
REDCombSent (Wu and Yu, 2014)	-	0.356	-	-
SimpBLEU-Recall (Song et al., 2013)	0.215	-	0.923	-
Depref-Align (Wu et al., 2013)	0.238	-	0.926	-
BLEU (Papineni et al., 2002)	0.197	0.285	0.854	0.888
Meteor (Denkowski and Lavie, 2014)	0.264	0.354	0.950	0.829
UPF-Cobalt	0.273	0.367	0.956	0.944

Table 1: Evaluation results on WMT13 and WMT14 datasets at segment and system levels

across different datasets with no need for parameter optimization.

In addition to the overall evaluation, we performed a series of ablation tests in order to assess the impact of the individual features of UPF-Cobalt. Each row in Table 2 below shows a feature excluded from the metric and the averaged Kendall's tau segment-level correlation for WMT14 dataset.

	Kendall's (τ)
UPF-Cobalt	0.367
(-) context penalty	0.319
(-) distrib. similarity	0.357
(-) weights on dep. functions	0.360
(-) equiv. dep. types	0.363

Table 2: Ablation test results

Context penalty. To estimate the benefit of using our context penalty we substituted it with fragmentation penalty from Meteor, which explicitly penalizes differences in sequential word order. As expected, this results in a significant drop in the correlation. Thus, this new component is indeed crucial for our metric's performance.

MWA has been shown to outperform Meteor in the alignment task. However, contrary to our expectations, simply using a more accurate aligner does not suffice to improve the correlation (Meteor achieves 0.354 correlation on this dataset). Manual inspection of the results shows that this is primarily due to the fact that MWA does not support phrase-level alignments. This functionality is highly relevant for the evaluation task as it allows covering acceptable variation that involves multi-word expressions. We plan to integrate phrasal alignments in the metric in the future.

Distributional similarity. Removing this component implies a considerable decrease in the correlation. Qualitative analysis of the results shows that its main contribution concerns cases of quasi-synonyms, i.e. words that can be considered synonymous only given the similarity of their contexts. The noise introduced by the component is neutralized by context penalty. If unrelated words are aligned, their context penalty will be high and aligning them won't increase sentence-level evaluation score. Also, in the ranking formulation of the evaluation task, distributional similarity helps to discriminate between low-quality translations. That is to say, it allows distinguishing sentences where words are at least minimally related from sentences, in which, for instance, source-language words are simply left untranslated.

Dependency weights. To test if giving different weights to contextual differences according to the dependency functions of the words involved, we put the values of all the weights to 1. This negatively affects the results, confirming that some differences are stronger indicators of MT errors than others. Thus, using the proposed weighting scheme the metric is capable of discriminating more or less serious MT errors based on the relative importance of mistranslated material.

Equivalence of syntactic constructions. Eliminating this functionality produces a smaller decrease in the correlation. Representing syntactic context as immediate neighbors of the word in a dependency graph allows covering a limited set of equivalent constructions, which are not frequent enough to have a significant impact on the results. The framework is flexible and more complex context equivalence definitions can be integrated in the future.

To appreciate the advantages of the metric, Table 3 provides a qualitative comparison of UPF-Cobalt's performance with strong baseline metric Meteor.[7] In this example, Meteor assigns low

[7] Stanford typed dependencies from Marneffe and Manning, (2008) are used for the description of syntactic relations.

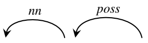
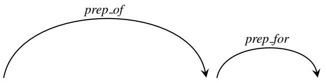
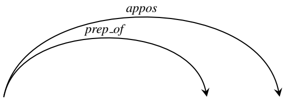
 Ref: An Obama voter 's cry of despair.	Equivalent dep. types	Scores	
		UPF-Cobalt	Meteor
 Cand1: The cry of despair of a voter for Obama.	*prep_of* ≈ *poss* *prep_for* ≈ *nn*	0.804	0.389
 Cand2: The cry of despair of a voter Obama.	*prep_of* ≈ *poss* *appos* ≠ *nn*	0.646	0.393

Table 3: Example of candidate and reference translations with the corresponding Meteor and UPF-Cobalt scores

scores to both candidate translations, due to the differences in word order and the presence of function words absent in the reference. However, it is clear that Candidate 1 is perfectly acceptable, whereas Candidate 2 contains an error concerning the relation between the words "voter" and "Obama". UPF-Cobalt correctly assigns a higher score to Candidate 1. Here all the content words are aligned and no context penalty is applied, since the syntactic contexts in which the words occur are equal or equivalent. Thus, *prep_for* relation in the candidate translation is equivalent to noun compound modifier relation *nn* in the reference and *prep_of* label in the candidate corresponds to possession modifier *poss* in the reference. UPF-Cobalt assigns a lower score to Candidate 2 due to the differences in the syntactic contexts of the words "voter" (context penalty – 0.426) and "Obama" (context penalty – 0.286), which constitute a translation error. Thus, context penalty values calculated for each pair of aligned words can be used for spotting and locating translation errors.

Qualitative analysis of the results also shows an interesting pattern in cases where UPF-Cobalt is outperformed by other metrics. This pattern is particularly relevant in the ranking evaluation setting. Consider the following example.

Ref: *Nevada has already completed a pilot.*
Cand1: *Nevada already has completed the pilot project.*
Cand2: *Nevada has already completed the pilot project.*

When ranking translations humans intend to avoid ties whenever possible. Both Candidate 1 and Candidate 2 are essentially correct, but the second translation is more adequate with regards to the norms and conventions of target language use. UPF-Cobalt assigns equal scores to both MTs. Thus, it successfully avoids penalizing acceptable differences in word order (the differences that do not affect the output of the dependency parser). However, it is not able to make more fine-grained distinctions regarding the fluency of MT. This issue can be addressed by integrating target language model features in the metric.

4 Related Work

Metrics based on string-level comparison take context into account in a simplistic manner. For instance, BLEU (Papineni et al., 2002) uses n-grams with length (1-4) and Meteor (Denkowski and Lavie, 2014) addresses the differences in sequential word order by means of fragmentation penalty, based on the number of adjacent aligned words. This often leads to penalizing acceptable differences induced by the use of semantically equivalent expressions. At the same time, spurious matches of the words that coincide in their surface form but play totally different roles in the corresponding sentences can incorrectly increase evaluation score.

To address these limitations a series of linguistically informed approaches have been proposed. Amigó et al. (2006) measure the degree of overlap between the dependency trees of candidate and reference translations. Giménez and Màrquez (2010) propose a combination of specialized similarity measures operating at different linguistic levels (lexical, syntactic and semantic). Guzman et al. (2014) further enrich this metric set with discourse level information. Padó et al. (2009) measure MT quality based on a rich set of features motivated by textual entailment.

Our work follows this line of research and exploits syntactic context to characterize the correspondences between the words in candidate and reference translations. In addition, we address the problem of syntactic variation that has rarely been dealt with in linguistically-informed MT evaluation. As shown in Fomicheva et al. (2015), this kind of variation is a regular source of differences between human reference and MT. Structural shifts (Ahrenberg and Merkel, 2000) are common practice in HT. Translators often introduce optional changes to the original sentence in order to adhere to specific principles of target language use, including stylistic issues and discourse processing conditions. MT may not contain such shifts but still be grammatically well-formed and perfectly deliver the contents of the source sentence. By taking into consideration the equivalence of syntactic constructions it is possible to avoid penalizing MT in these cases.

5 Conclusions and Future Work

We have shown that using contextual information helps to distinguish candidate translations that are different from the reference and still essentially correct from those that share high number of words with HT but fail to preserve the meaning of the source sentence due to translation errors.

Also, we enhanced existing methods for addressing meaning-preserving variation by exploiting distributional similarity at lexical level and classes of equivalent dependency types at syntactic level. The results demonstrate that the metric achieves competitive performance on WMT13 and WMT14 data.

As future work, we consider improving the metric by extending the alignment component to phrase-level and refining the equivalent dependency types to increase the coverage of linguistic variation at syntactic level. Another interesting direction would be to integrate target-language features and take into consideration the properties of non-aligned material. Finally, we plan to test if the metric can be successfully used for error detection and classification.

References

Lars Ahrenberg and Magnus Merkel. 2000. Correspondence Measures for MT Evaluation. In *Proceedings of the Second International Conference on Linguistic Resources and Evaluation*, 1255–1261. Athens, Greece.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Chrstof Monz, and Josh Schroeder. 2007. (Meta-)Evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, 136–158. Prague, Czech Republic. Association for Computational Linguistics (ACL).

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, USA. ACL.

Marina Fomicheva, Núria Bel, and Iria da Cunha. 2015. Neutralizing the Effect of Translation Shifts on Automatic Machine Translation Evaluation. In *Gelbukh, Alexander (ed.) Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015: Proceedings 1*, 596–607.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 758–764. ACL.

Jesus Giménez and Lluís Màrquez. 2010. Linguistic Measures for Automatic Machine Translation Evaluation. *Machine Translation*, 24(3-4):77–86.

Francisco Guzman, Shafiq Joty, Lluís Màrquez, Alessandro Moschitti, Preslav Nakov, and Massimo Nicosia. 2014. Learning to Differentiate Better from Worse Translations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 214–220. Doha, Qatar. ACL.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the ACL (Volume 2: Short Papers)*. Baltimore, USA. ACL.

Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A Phrase-Based Alignment Model for Natural Language Inference. In *Proceedings of the 2008 EMNLP Conference*, 214–220. Honolulu, USA. ACL.

Bill MacCartney, Trond Grenager, Marie de Marneffe, Daniel Cer, and Christopher D. Manning. 2006. Learning to recognize features of valid textual entailments. In *Proceedings of the NAACL – Human Language Technologies*.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, USA. ACL.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 45–51. Sofia, Bulgaria. ACL.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford typed dependencies manual. Technical Report, Stanford University.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the International Conference on Language Resources and Evaluation*, 449–454.

George Miller and Christiane Fellbaum. 2007. WordNet. `http://wordnet.princeton.edu`.

Sebastian Padó, Daniel Cer, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Measuring machine translation quality as semantic equivalence: A metric based on entailment features. *Machine Translation*, 23:181–193.

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the ACL*, 311–318. Philadelphia, USA.

Xingyi Song, Trevor Cohn, and Lucia Specia. 2013. BLEU deconstructed: Designing a better MT Evaluation Metric. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*.

Miloš Stanojević and Khalil Sima'an. 2014. BEER: A Smooth Sentence Level Evaluation Metric with Rich Ingredients. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, USA. ACL.

Arafat Md Sultan, Steven Bethard, and Tamara Summer. 2014. Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence. *Transactions of the Association of Computational Linguistics*, volume 2(1):219–230.

Kapil Thadani, Scott Martin, and Michael White. 2012. A Joint Phrasal and Dependency Model for Paraphrase Alignment. In *Proceedings of 24th International Conference on Computational Linguistics, COLING 2012*, 1229–1238. Bombay, India.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Xiaofeng Wu, Hui Yu, and Qun Liu. 2014. RED: DCU-CASICT Participation in WMT2014 Metrics Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, USA. ACL.

Xiaofeng Wu, Hui Yu, and Qun Liu. 2013. DCU Participation in WMT2013 Metrics Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria. ACL.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. Semi-Markov Phrase-based Monolingual Alignment. In *Proceedings of the 2013 EMNLP Conference*, 590–600. ACL.