CobaltF: A Fluent Metric for MT Evaluation

Marina Fomicheva, Núria Bel
IULA, Universitat Pompeu Fabra
firstname.lastname@upf.edu

Lucia Specia
University of Sheffield, UK
l.specia@sheffield.ac.uk

Iria da Cunha
Univ. Nacional de Educación a Distancia
iriad@flog.uned.es

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

Abstract

The vast majority of Machine Translation (MT) evaluation approaches are based on the idea that the closer the MT output is to a human reference translation, the higher its quality. While translation quality has two important aspects, adequacy and fluency, the existing reference-based metrics are largely focused on the former. In this work we combine our metric UPF-Cobalt, originally presented at the WMT15 Metrics Task, with a number of features intended to capture translation fluency. Experiments show that the integration of fluency-oriented features significantly improves the results, rivalling the best-performing evaluation metrics on the WMT15 data.

1 Introduction

Automatic evaluation plays an instrumental role in the development of Machine Translation (MT) systems. It is aimed at providing fast, inexpensive, and objective numerical measurements of translation quality. As a cost-effective alternative to manual evaluation, the main concern of automatic evaluation metrics is to accurately approximate human judgments.

The vast majority of evaluation metrics are based on the idea that the closer the MT output is to a human reference translation, the higher its quality. The evaluation task, therefore, is typically approached by measuring some kind of similarity between the MT (also called candidate translation) and a reference translation. The most widely used evaluation metrics, such as BLEU (Papineni et al., 2002), follow a simple strategy of counting the number of matching words or word sequences in the candidate and reference translations. Despite its wide use and practical utility, automatic evaluation based on a straightforward candidate-reference comparison has long been criticized for its low correlation with human judgments at sentence-level (Callison-Burch and Osborne, 2006).

The core aspects of translation quality are fidelity to the source text (or adequacy, in MT parlance) and acceptability (also termed fluency) regarding the target language norms and conventions (Toury, 2012). Depending on the purpose and intended use of the MT, manual evaluation can be performed in a number of different ways. However, in any setting both adequacy and fluency shape human perception of the overall translation quality.

By contrast, automatic reference-based metrics are largely focused on MT adequacy, as they do not evaluate the appropriateness of the translation in the context of the target language. Translation fluency is thus assessed only indirectly, through the comparison with the reference. However, the difference from a particular human translation does not imply that the MT output is disfluent (Fomicheva et al., 2015a).

We propose to explicitly model translation fluency in reference-based MT evaluation. To this end, we develop a number of features representing translation fluency and integrate them with our reference-based metric UPF-Cobalt, which was originally presented at WMT15 (Fomicheva et al., 2015b). Along with the features based on the target Language Model (LM) probability of the MT output, which have been widely used in the related fields of speech recognition (Uhrik and Ward, 1997) and quality estimation (Specia et al., 2009), we design a more detailed representation of MT fluency that takes into account the number of disfluent segments observed in the candidate translation. We test our approach with the data avail-

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 483–490,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

able from WMT15 Metrics Task and obtain very promising results, which rival the best-performing system submissions. We have also submitted the metric to the WMT16 Metrics Task.

2 Related Work

The recent advances in the field of MT evaluation have been largely directed to improving the informativeness and accuracy of candidate-reference comparison. Meteor (Denkowski and Lavie, 2014) allows for stem, synonym and paraphrase matches, thus addressing the problem of acceptable linguistic variation at lexical level. Other metrics measure syntactic (Liu and Gildea, 2005), semantic (Lo et al., 2012) or even discourse similarity (Guzmán et al., 2014) between candidate and reference translations. Further improvements have been recently achieved by combining these partial measurements using different strategies including machine learning techniques (Comelles et al., 2012; Giménez and Màrquez, 2010b; Guzmán et al., 2014; Yu et al., 2015). However, none of the above approaches explicitly addresses the fluency of the MT output.

Predicting MT quality with respect to the target language norms has been investigated in a different evaluation scenario, when human translations are not available as benchmark. This task, referred to as confidence or quality estimation, is aimed at MT systems in use and therefore has no access to reference translations (Specia et al., 2010).

Quality estimation can be performed at different levels of granularity. Sentence-level quality estimation (Specia et al., 2009; Blatz et al., 2004) is addressed as a supervised machine learning task using a variety of algorithms to induce models from examples of MT sentences annotated with quality labels. In the word-level variant of this task, each word in the MT output is to be judged as correct or incorrect (Luong et al., 2015; Bach et al., 2011), or labelled for a specific error type.

Research in the field of quality estimation is focused on the design of features and the selection of appropriate learning schemes to predict translation quality, using source sentences, MT outputs, internal MT system information and source and target language corpora. In particular, features that measure the probability of the MT output with respect to a target LM, thus capturing translation fluency, have demonstrated highly competitive performance in a variety of settings (Shah et al., 2013).

Both translation evaluation and quality estimation aim to evaluate MT quality. Surprisingly, there have been very few attempts at joining the insights from these two related tasks. A notable exception is the work by Specia and Giménez (2010), who explore the combination of a large set of quality estimation features extracted from the source sentence and the candidate translation, as well as the source-candidate alignment information, with a set of 52 MT evaluation metrics from the `Asiya Toolkit` (Giménez and Màrquez, 2010a). They report a significant improvement over the reference-based evaluation systems on the task of predicting human post-editing effort. We follow this line of research by focusing specifically on integrating fluency information into reference-based evaluation.

3 UPF-Cobalt Review

UPF-Cobalt[1] is an alignment-based evaluation metric. Following the strategy introduced by the well known Meteor (Denkowski and Lavie, 2014), UPF-Cobalt's score is based on the number of aligned words with different levels of lexical similarity. The most important feature of the metric is a syntactically informed context penalty aimed at penalizing the matches of similar words that play different roles in the candidate and reference sentences. The metric has achieved highly competitive results on the data from previous WMT tasks, showing that the context penalty allows to better discriminate between acceptable candidate-reference differences and the differences incurred by MT errors (Fomicheva et al., 2015b). Below we briefly review the main components of the metric. For a detailed description of the metric the reader is referred to (Fomicheva and Bel, 2016).

3.1 Alignment

The alignment module of UPF-Cobalt builds on an existing system – Monolingual Word Aligner (MWA), which has been shown to significantly outperform state-of-the-art results for monolingual alignment (Sultan et al., 2014). We increase the coverage of the aligner by comparing distributed word representations as an additional source of lexical similarity information,

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

which allows to detect cases of quasi-synonyms (Fomicheva and Bel, 2016).

3.2 Scoring

UPF-Cobalt's sentence-level score is a weighted combination of precision and recall over the sum of the individual scores computed for each pair of aligned words. The word-level score for a pair of aligned words (t, r) in the candidate and reference translations is based on their lexical similarity $(LexSim)$ and a context penalty which measures the difference in their syntactic contexts (CP):

$$score(t,r) = LexSim(t,r) - CP(t,r)$$

Lexical similarity is defined based on the type of lexical match (exact match, stem match, synonyms, etc.)[2] (Denkowski and Lavie, 2014). The crucial component of the metric is the context penalty, which is applied at word-level to identify the cases where the words are aligned (i.e. lexically similar) but play different roles in the candidate and reference translations and therefore should contribute less to the sentence-level 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. The context penalty (CP) is computed as follows:

$$CP(t,r) = \frac{\sum_{1..i} w(C_i^*)}{\sum_{1..i} w(C_i)} \times ln \left(\sum_{1..i} w(C_i) + 1 \right)$$

where w refers to the weights that reflect the relative importance of the dependency functions of the context words, 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 context penalty is calculated for both candidate and reference words. The metric computes an average between reference-side context penalty and candidate-side context penalty for each word pair. The sentence-level average can be obtained in a straightforward way from the word-level values (we use it as a feature in the decomposed version of the metric below).

4 Approach

In this paper we learn an evaluation metric that combines a series of adequacy-oriented features extracted from the reference-based metric UPF-Cobalt with various features intended to focus on translation fluency. This section first describes the metric-based features used in our experiments and then the selection and design of our fluency-oriented features.

4.1 Adequacy-oriented Features

UPF-Cobalt incorporates in a single score various distinct MT characteristics (lexical choice, word order, grammar issues, such as wrong word forms or wrong choice of function words, etc.). We note that these components can be related, to a certain extent, to the aspects of translation quality being discussed in this paper. The syntactic context penalty of UPF-Cobalt is affected by the well-formedness of the MT output, and may reflect, although indirectly, grammaticality and fluency, whereas the proportion of aligned words depends on the correct lexical choice.

Using the components of the metric instead of the scores yields a more fine-grained representation of the MT output. We explore this idea in our experiments by designing a decomposed version of UPF-Cobalt. More specifically, we use 48 features (grouped below for space reasons):

- Percentage and number of aligned words in the candidate and reference translations
- Percentage and number of aligned words with different levels of lexical similarity in the candidate and reference translations
- Percentage and number of aligned function and content words in the candidate and reference translations
- Minimum, maximum and average context penalty
- Percentage and number of words with high context penalty[4]
- Number of words in the candidate and reference translations

[2] Specifically, the values for different types of lexical similarity are: same word forms - 1.0, lemmatizing or stemming - 0.9, WordNet synsets - 0.8, paraphrase database - 0.6 and distributional similarity - 0.5.

[3] The weights w are: argument/complement functions - 1.0, modifier functions - 0.8 and specifier/auxiliary functions - 0.2.

[4] These are words with the context penalty value higher than the average computed on the training set used in our experiments.

4.2 Fluency-oriented Features

We suggest that the fluency aspect of translation quality has been overlooked in the reference-based MT evaluation. Even though syntactically-informed metrics capture structural differences and are, therefore, assumed to account for grammatical errors, we note that the distinction between adequacy and fluency is not limited to grammatical issues and thus exists at all linguistic levels. For instance, at lexical level, the choice of a particular word or expression may be similar in meaning to the one present in the reference (adequacy), but awkward or even erroneous if considered in the context of the norms of the target language use. Conversely, due to the variability of linguistic expression, neither lexical nor syntactic differences from a particular human translation imply ill-formedness of the MT output.

Sentence fluency can be described in terms of the frequencies of the words with respect to a target LM. Here, in addition to the LM-based features that have been shown to perform well for sentence-level quality estimation (Shah et al., 2013), we introduce more complex features derived from word-level n-gram statistics. Besides the word-based representation, we rely on Part-of-Speech (PoS) tags. As suggested by (Felice and Specia, 2012), morphosyntactic information can be a good indicator of ill-formedness in MT outputs.

First, we select 16 simple sentence-level features from previous work (Felice and Specia, 2012; Specia et al., 2010), summarized below.

- Number of words in the candidate translation
- LM probability and perplexity of the candidate translation
- LM probability of the candidate translation with respect to an LM trained on a corpus of PoS tags of words
- Percentage and number of content/function words
- Percentage and number of verbs, nouns and adjectives

Essentially, these features average LM probabilities of the words to obtain a sentence-level measurement. While being indeed predictive of sentence-level translation fluency, they are not representative of the number and scale of the disfluent fragments contained in the MT sentence. Moreover, if an ill-formed translation contains various word combinations that have very high probability according to the LM, the overall sentence-level LM score may be misleading.

To overcome the above limitations, we use word-level n-gram frequency measurements and design various features to extend them to the sentence level in a more informative way. We rely on LM backoff behaviour, as defined in (Raybaud et al., 2011). LM backoff behaviour is a score assigned to the word according to how many times the target LM had to back-off in order to assign a probability to the word sequence. The intuition behind is that an n-gram not found in the LM can indicate a translation error. Specifically, the backoff behaviour value $b(w_i)$ for a word w_i in position i of a sentence is defined as:

$$b(w_i) = \begin{cases} 7, & \text{if } w_{i-2}, w_{i-1}, w_i \text{ exists in the model} \\ 6, & \text{if } w_{i-2}, w_{i-1} \text{and } w_{i-1}, w_i \text{ both exist} \\ & \text{ in the model} \\ 5, & \text{if only } w_{i-1}, w_i \text{ exists in the model} \\ 4, & \text{if only } w_{i-2}, w_{i-1} \text{and } w_i \text{ exist} \\ & \text{separately in the model} \\ 3, & \text{if } w_{i-1} \text{and } w_i \text{ both exist} \\ & \text{in the model} \\ 2, & \text{if only } w_i \text{ exists in the model} \\ 1, & \text{if } w_i \text{ is an out-of-vocabulary word} \end{cases}$$

We compute this score for each word in the MT output and then use the mean, median, mode, minimum and maximum of the backoff behaviour values as separate sentence-level features. Also, we calculate the percentage and number of words with low backoff behaviour values (< 5) to approximate the number of fluency errors in the MT output.

Furthermore, we introduce a separate feature that counts the words with a backoff behaviour value of 1, i.e. the number of out-of-vocabulary (OOV) words. OOV words are indicative of the cases when source words are left untranslated in the MT. Intuitively, this should be a strong indicator of low MT quality.

Finally, we note that UPF-Cobalt, not unlike the majority of reference-based metrics, lacks information regarding the MT words that are not aligned or matched to any reference word. Such fragments do not necessarily constitute an MT error, but may be due to acceptable linguistic variations. Collecting fluency information specifically for these fragments may help to distinguish acceptable variation from MT errors. If a candidate word or phrase is absent from the reference but is fluent in the target language, then the difference is possibly not indicative of an error and should be penalized less. Based on this observation, we introduce a separate set of features that compute the word-level measurements discussed above only for the words that are not aligned to the reference translation.

This results in 49 additional features, grouped here for space reasons:

- Summary statistics of the LM backoff behaviour (word and PoS-tag LM)
- Summary statistics of the LM backoff behaviour for non-aligned words only (word and PoS tag LM)
- Percentage and number of words with low backoff behaviour value (word and PoS tag LM)
- Percentage and number of non-aligned words with low backoff behaviour value (word and PoS tag LM)
- Percentage and number of OOV words
- Percentage and number of non-aligned OOV words

5 Experimental Setup

For our experiments, we use the data available from the WMT14 and WMT15 Metrics Tasks for into-English translation directions. The datasets consist of source texts, human reference translations and the outputs from the participating MT systems for different language pairs. During manual evaluation, for each source sentence the annotators are presented with its human translation and the outputs of a random sample of five MT systems, and asked to rank the MT outputs from best to worst (ties are allowed). Pairwise system comparisons are then obtained from this compact annotation. Details on the WMT data for each language pair are given in Table 1.

	WMT14			WMT15		
LP	**Rank**	**Sys**	**Src**	**Rank**	**Sys**	**Src**
Cs-En	21,130	5	3,003	85,877	16	2,656
De-En	25,260	13	3,003	40,535	13	2,169
Fr-En	26,090	8	3,003	29,770	7	1,500
Ru-En	34,460	13	3,003	44,539	13	2,818
Hi-En	20,900	9	2,507	-	-	-
Fi-En	-	-	-	31,577	14	1,370

Table 1: Number of pairwise comparisons (Rank), translation systems (Sys) and source sentences (Src) per language pair for the WMT14 and WMT15 datasets

In our work we focus on sentence-level metrics' performance, which is assessed by converting metrics' scores to ranks and comparing them to the human judgements with Kendall rank correlation coefficient (τ). We use the WMT14 official Kendall's Tau implementation (Macháček and Bojar, 2014). Following the standard practice at WMT and to make our work comparable to the official metrics submitted to the task, we exclude ties in human judgments both for training and for testing our system.

Our model is a simple linear interpolation of the features presented in the previous sections. For tuning the weights, we use the learn-to-rank approach (Burges et al., 2005), which has been successfully applied in similar settings in previous work (Guzmán et al., 2014; Stanojevic and Sima'an, 2015). We use a standard implementation of Logistic Regression algorithm from the Python toolkit `scikit-learn`[5]. The model is trained on WMT14 dataset and tested on WMT15 dataset.

For the extraction of word-level backoff behaviour values and sentence-level fluency features, we use `Quest++`[6], an open source tool for quality estimation (Specia et al., 2015). We employ the LM used to build the baseline system for WMT15 Quality Estimation Task (Bojar et al., 2015).[7] This LM provided was trained on data from the WMT12 translation task (a combination of news and Europarl data) and thus matches the domain of the dataset we use in our experiments. PoS tagging was performed with TreeTagger (Schmid, 1999).

6 Experimental Results

Table 2 summarizes the results of our experiments. Group I presents the results achieved by UPF-Cobalt and its decomposed version described in Section 4.1. Contrary to our expectations, the performance is slightly degraded when using the metrics' components (UPF-Cobalt$_{comp}$). Our intuition is that this happens due to the sparseness of the features based on the counts of different types of lexical matches.

Group II reports the performance of the fluency features presented in Section 4.2. First of all, we note that these features on their own (FeaturesF)

[5] `http://scikit-learn.org/`

[6] `https://github.com/ghpaetzold/questplusplus`

[7] `http://www.statmt.org/wmt15/quality-estimation-task.html.`

	Metric	cs-en	de-en	fi-en	fr-en	ru-en	Avg τ
I	UPF-Cobalt	.457±.011	.427±.011	.437±.011	.386±.011	.402±.011	.422±.011
	UPF-Cobalt$_{comp}$	.442±.011	.418±.011	.428±.011	.387±.011	.388±.011	.413±.012
II	FeaturesF	.373±.011	.337±.011	.359±.011	.267±.011	.263±.011	.320±.011
	CobaltF$_{simple}$	.487±.011	.445±.011	.455±.011	.401±.011	.395±.011	.437±.012
	CobaltF$_{comp}$	.481±.011	.438±.011	.464±.011	.403±.011	.395±.011	.436±.011
	MetricsF	.502±.011	.457±.011	.450±.011	.413±.011	.410±.011	**.447±.011**
III	DPMFcomb	.495±.011	.482±.011	.445±.011	.395±.011	.418±.011	.447±.011
	BEER_Treepel	.471±.011	.447±.011	.438±.011	.389±.011	.403±.011	.429±.011
	RATATOUILLE	.472±.011	.441±.011	.421±.011	.398±.011	.393±.011	.425±.010
IV	BLEU	.391±.011	.360±.011	.308±.011	.358±.011	.329±.011	.349±.011
	Meteor	.439±.011	.422±.011	.406±.011	.380±.011	.386±.011	.407±.012

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

achieve a reasonable correlation with human judgments, showing that fluency information is often sufficient to compare the quality of two candidate translations. Secondly, fluency features yield a significant improvement when used together with the metrics' score (CobaltF$_{simple}$) or with the components of the metric (CobaltF$_{comp}$). We further boost the performance by combining the scores of the metrics BLEU, Meteor and UPF-Cobalt with our fluency features (MetricsF).

The results demonstrate that fluency features provide useful information regarding the overall translation quality, which is not fully captured by the standard candidate-reference comparison. These features are discriminative when the relationship to the reference does not provide enough information to distinguish between the quality of two alternative candidate translations. For example, it may well be the case that both MT outputs are very different from human reference, but one constitutes a valid alternative translation, while the other is totally unacceptable.

Finally, Groups III and VI contain the results of the best-performing evaluation systems from the WMT15 Metrics Task, as well as the baseline BLEU metric (Papineni et al., 2002) and a strong competitor, Meteor (Denkowski and Lavie, 2014), which we reproduce here for the sake of comparison. DPMFComb (Yu et al., 2015) and RATATOUILLE (Marie and Apidianaki, 2015) use a learnt combination of the scores from different evaluation metrics, while BEER_Treepel (Stanojevic and Sima'an, 2015) combines word matching, word order and syntax-level features. We note that the number and complexity of the metrics used in the above approaches is quite high. For instance, DPMFComb is based on 72 separate evaluation systems, including the resource-heavy linguistic metrics from the `Asiya Toolkit` (Giménez and Màrquez, 2010a).

7 Conclusions

The performance of reference-based MT evaluation metrics is limited by the fact that dissimilarities from a particular human translation do not always indicate bad MT quality. In this paper we proposed to amend this issue by integrating translation fluency in the evaluation. This aspect determines how well a translated text conforms to the linguistic regularities of the target language and constitutes a strong predictor of the overall MT quality.

In addition to the LM-based features developed in the field of quality estimation, we designed a more fine-grained representation of translation fluency, which in combination with our reference-based evaluation metric UPF-Cobalt yields a highly competitive performance for the prediction of pairwise preference judgments. The results of our experiments thus confirm that the integration of features intended to address translation fluency improves reference-based MT evaluation.

In the future we plan to investigate the performance of fluency features for the modelling of other types of manual evaluation, such as absolute scoring.

Acknowledgments

This work was partially funded by TUNER (TIN2015-65308-C5-5-R) and MINECO/FEDER, UE. Marina Fomicheva was supported by funding from the FI-DGR grant program of the Generalitat de Catalunya. Iria da Cunha was supported by a Ramón y Cajal contract (RYC-2014-16935). Lucia Specia was supported by the QT21 project (H2020 No. 645452).

References

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: A Method for Measuring Machine Translation Confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 211–219. Association for Computational Linguistics (ACL).

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence Estimation for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321. ACL.

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

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to Rank Using Gradient Descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96. ACM.

Chris Callison-Burch and Miles Osborne. 2006. Reevaluating the Role of BLEU in Machine Translation Research. In *In Proceedings of the European Association for Computational Linguistics (EACL)*, pages 249–256. ACL.

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

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*, pages 376–380.

Mariano Felice and Lucia Specia. 2012. Linguistic Features for Quality Estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 96–103. ACL.

Marina Fomicheva and Núria Bel. 2016. Using Contextual Information for Machine Translation Evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2755–2761.

Marina Fomicheva, Núria Bel, and Iria da Cunha. 2015a. Neutralizing the Effect of Translation Shifts on Automatic Machine Translation Evaluation. In *Computational Linguistics and Intelligent Text Processing*, pages 596–607.

Marina Fomicheva, Núria Bel, Iria da Cunha, and Anton Malinovskiy. 2015b. UPF-Cobalt Submission to WMT15 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 373–379.

Jesús Giménez and Lluís Màrquez. 2010a. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Jesús Giménez and Lluís Màrquez. 2010b. Linguistic Measures for Automatic Machine Translation Evaluation. *Machine Translation*, 24(3):209–240.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using Discourse Structure Improves Machine Translation Evaluation. In *ACL (1)*, pages 687–698.

Ding Liu and Daniel Gildea. 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.

Chi-Kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. 2012. Fully Automatic Semantic MT Evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 243–252. ACL.

Ngoc-Quang Luong, Laurent Besacier, and Benjamin Lecouteux. 2015. Towards Accurate Predictors of Word Quality for Machine Translation: Lessons Learned on French–English and English–Spanish Systems. *Data & Knowledge Engineering*, 96:32–42.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 Metrics Shared Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301.

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

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 ACL*, pages 311–318. ACL.

Sylvain Raybaud, David Langlois, and Kamel Smaïli. 2011. this sentence is wrong. detecting errors in machine-translated sentences. *Machine Translation*, 25(1):1–34.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Kashif Shah, Trevor Cohn, and Lucia Specia. 2013. An Investigation on the Effectiveness of Features for Translation Quality Estimation. In *Proceedings of the Machine Translation Summit*, volume 14, pages 167–174.

Lucia Specia and Jesús Giménez. 2010. Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation. In *The Ninth Conference of the Association for Machine Translation in the Americas*.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the Sentence-level Quality of Machine Translation Systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine Translation Evaluation versus Quality Estimation. *Machine Translation*, 24(1):39–50.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level Translation Quality Prediction with QuEst++. In *53rd Annual Meeting of the Association for Computational Linguistics and Seventh International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: System Demonstrations*, pages 115–120.

Miloš Stanojevic and Khalil Sima'an. 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.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence. *Transactions of the ACL*, 2:219–230.

Gideon Toury. 2012. *Descriptive Translation Studies and beyond: Revised edition*, volume 100. John Benjamins Publishing.

C. Uhrik and W. Ward. 1997. Confidence Metrics Based on N-gram Language Model Backoff Behaviors. In *Proceedings of Fifth European Conference on Speech Communication and Technology*, pages 2771–2774.

Hui Yu, Qingsong Ma, Xiaofeng Wu, and Qun Liu. 2015. CASICT-DCU Participation in WMT2015 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 417–421.