UHH Submission to the WMT17 Metrics Shared Task

Melania Duma and Wolfgang Menzel
University of Hamburg
Natural Language Systems Division
{duma, menzel}@informatik.uni-hamburg.de

Abstract

In this paper the UHH submission to the WMT17 Metrics Shared Task is presented, which is based on sequence and tree kernel functions applied to the reference and candidate translations. In addition we also explore the effect of applying the kernel functions on the source sentence and a back-translation of the MT output, but also on the pair composed of the candidate translation and a pseudo-reference of the source segment. The newly proposed metric was evaluated using the data from WMT16, with the results demonstrating a high correlation with human judgments.

1 Introduction

The evaluation of Machine Translation (MT) represents a very important domain of research, as providing meaningful, automatic and accurate methods for determining the quality of machine-translated output is a key component in the development cycle of a MT system. However, the task is inherently difficult due to the expressiveness of natural language, which often allows conveying a message in more than one equivalent ways. When translating from a source language into a target one, the input data for evaluation conventionally consists of a set of tuples, with each tuple composed of:

- a source segment, representing the sentence to be translated in the source language
- a candidate translation (also known as a target segment), obtained by translating the source segment into the target language using an MT system
- a reference translation, representing a correct human-generated translation of the source segment

As a research field, MT evaluation can be divided into two categories: reference-free evaluation and reference-based one. The reference-free evaluation, also known as Quality Estimation, aims at providing automatic methods, for assessing the quality of candidate translations, which do not require reference translations. In the case of a reference-based evaluation, the target segment is compared with the reference translation resulting in a score that measures the similarity between the two sentences. Different approaches for computing the comparison have been implemented, with the most frequently used one being BLEU (Papineni et al., 2002), which measures the quality of the candidate translation by counting the number of n-grams it has in common with the reference translations. Nonetheless, multiple disadvantages of BLEU have already been pointed out, as in Callison-Burch et al. (2006), where it is shown that an increase of the BLEU score does not necessary correlate with a better performing system. This has motivated further research into additional MT evaluation methods that rely on more than lexical matching by additionally including the syntactic and semantic structure of the sentences (e.g. (Popović and Ney, 2009), (Gautam and Bhattacharyya, 2014)).

We propose a new method for the evaluation of MT output, based on tree and sequence kernel functions, applied on the pair of reference and candidate translations. In addition, we study the impact of applying the kernels on the tuple consisting of the source segment and a back-translation, together with the pair comprised of the candidate translation and a pseudo-reference. A pseudo-reference is the result of translating the source

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 582–588
Copenhagen, Denmark, September 711, 2017.

segment into the target language, while a back-translation is obtained by translating the target segment into the source language. The evaluation results show that the new metric strongly correlates with human judgments, outperforming the state-of-the-art methods.

2 Related work

MT evaluation methods can be categorized according to the level of analysis that they address into lexical ones (e.g BLEU (Papineni et al., 2002), TER (Snover et al., 2006)), syntactic ones (e.g. (Popović and Ney, 2007), (Giménez and Màrquez, 2007)) or semantic ones (e.g. (Castillo and Estrella, 2012)), with hybrid combinations integrating more than one representational layer at a time.

A syntactic evaluation method based on tree kernels is proposed in Liu and Gildea (2005). It uses the subtree kernel introduced in Collins and Duffy (2002) to calculate the similarity between the reference and the candidate translations. Besides this, a syntactic metric based on counting the number of fixed-depth subtrees shared by the two translations is also introduced, with both metrics being applied on the constituency trees of the input data. Additionally, a dependency tree based metric is presented, which computes the number of common headword chains, where a headword chain is defined as the concatenation of words that form a path in the dependency tree.

Another MT evaluation method that makes use of tree kernels is introduced in Guzmán et al. (2014). It also uses the subtree kernel introduced in Collins and Duffy (2001), but in this case it calculates the similarity between the discourse trees of the candidate and reference translation. The evaluation combined the newly proposed metric with already existing ones and the results showed that the addition is beneficial for improving the correlation scores.

The role of back-translations has also been investigated before, like in the case of Rapp (2009) where the quality of a candidate translation is assessed by measuring the similarity, in terms of a modified version of BLEU, between its back-translation and the initial source segment. In the case of pseudo-references, they have been used as an additional source of data for tuning the parameters of MT systems, like in the case of Ammar et al. (2013). An evaluation method based on pseudo-references is presented in Albrecht and Hwa (2007) and then further extended in Albrecht and Hwa (2008), where a metric is trained to correlate with human judgments based on features extracted with the help of three pseudo-references. The features are in the form of 18 kinds of reference-based scores together with an additional set of 25 monolingual fluency scores. The results showed that the new metric correlates well with human assessments and generalizes well across different language pairs.

The novelty of the MT Evaluation metric introduced in this paper is twofold. First of all, the method makes use of the Partial Tree Kernel (PTK), a more general type of kernel function, which to the authors' knowledge has not been applied in the context of MT metrics-based evaluation before. Secondly, the proposed method also explores what impact do sequence kernels (SK) have on the quality of a kernel evaluation metric, by studying its potential individually, but also in combination with the Partial Tree Kernel. Furthermore, we extend on the previous work of pseudo-references and back-translations by studying their impact in the context of using them as input data for kernel functions.

3 Methods and implementation

A kernel function makes use of structural representations of the input data in order to calculate the number of substructures they share, without explicitly stating the feature spaces corresponding to the two representations (Moschitti, 2006a). The types of representations taken into account can be, among others, vectorial, sequential or tree-based. The tree kernels developed so far distinguish themselves from one another by the types of tree fragments (e.g. subsets, subtrees or partial trees) and the type of syntactic trees (constituency or dependency) they employ in their computation, which influences their suitability for certain tasks (see Moschitti (2006a)). Contrastively, sequence kernels (e.g (Bunescu and Mooney, 2005), (Nguyen et al., 2009)) make use of subsequences in the computation of the kernel.

The new method for the evaluation of Machine Translation proposed in this paper, denoted as TSKM, makes use of both tree and sequence kernels, which are applied on the pair of candidate and reference translations. The tree kernel used is represented by the Partial Tree Kernel (PTK)

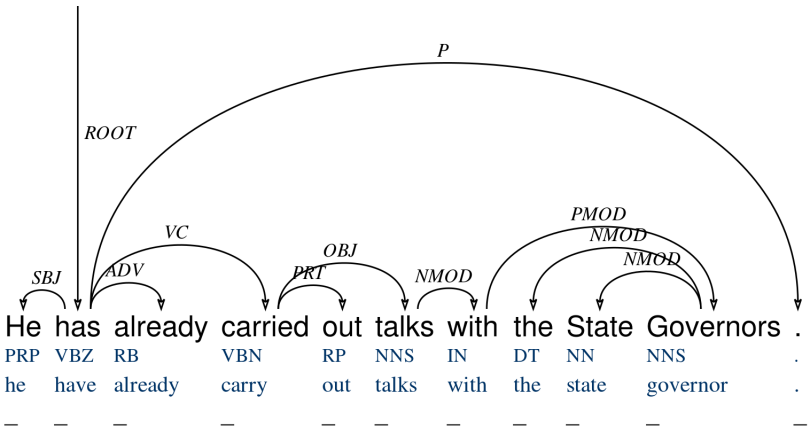

Figure 1: Example of a dependency tree.

(Moschitti, 2006a). It uses partial tree fragments, which are a generalization over subtrees and subset trees, so that a node and its partial descendants can constitute a valid fragment. An example of a dependency tree is presented in Figure 1 [1] and some possible partial trees for it are *(has(carried(out)))* or *(carried(out talks))*. For the sequence kernel (SK), the kernel introduced in Bunescu and Mooney (2005) is utilized, which computes the number of common patterns shared by the two input sentences.

Formally, TSKM can be defined as:

$$TSKM_{basic} = TSKM(r, c) = \frac{PTK(r,c)+SK(r,c)}{2} \quad (1)$$

with *r* and *c* denoting the reference and the candidate translations and *PTK* and *SK* referring to the scores of the Partial Tree Kernel and the Sequence Kernel.

Furthermore, we experimented with using an additional pseudo-reference and a back-translation in the computation of the metric in order to explore how the different combination schemes influence the performance of TSKM. One possible kind of combination can be represented as:

$$TSKM_{comb} = \frac{TSKM_{basic}+TSKM_{pseudo}+TSKM_{back}}{3} \quad (2)$$

$$TSKM_{pseudo} = TSKM(c, s_t) = \frac{PTK(c,s_t)+SK(c,s_t)}{2} \quad (3)$$

$$TSKM_{back} = TSKM(s, c_t) = \frac{PTK(s,c_t)+SK(s,c_t)}{2} \quad (4)$$

with s_t and c_t representing the pseudo-reference and the back-translation respectively and *s* marking the source segment. Our rationale for utilizing the pseudo-reference was motivated by two factors. In the first place, we wanted to determine whether an additional reference, even if only an approximate one, helps to better predict the quality of the candidate translation. Furthermore, we also wanted to investigate the possibility to apply our new evaluation method, in the scenario without official reference translation. Producing reference translations is a time-consuming and expensive task, therefore an evaluation method that performs well even without reference translations being available would be highly desirable.

In the case of back-translations, we wanted to investigate if the quality of a candidate translation can be approximated using the quality of its back-translation. This would prove extremely beneficial, especially in the case of low-resource language pairs, where no high quality analysis tools (e.g lemmatizers, pos-taggers or parsers) for the target language are available, a situation that would prevent TSKM from being applied. In our experiments, both the pseudo-references together with the back-translations were obtained using the free online Google Translator Toolkit [2].

To apply the tree and sequence kernels for the task of Machine Translation evaluation, a preprocessing of the input data is necessary. In the case of PTK, the input data was first tokenized and pos-tagged, followed by a parsing step using the Bohnet graph-based dependency parser (Bohnet, 2010) and the publicly available syntactic analy-

[1] The visualization was obtained using Arborator https://arborator.ilpga.fr/q.cgi (Gerdes, 2013)

[2] https://translate.google.com/toolkit

TSKM	not exact					exact				
	cs-en	**fi-en**	**ru-en**	**tr-en**	**Average**	**cs-en**	**fi-en**	**ru-en**	**tr-en**	**Average**
SK(r,c)	.997	.939	.958	.973	.967	.995	.850	.957	.924	.932
SK(c,s_t)	.993	.426	.948	.948	.829	.995	.391	.949	.895	.808
SK(r,c)+SK(c,s_t)	.998	.507	.955	.961	.855	.998	.454	.955	.908	.829
PTK(r,c)	.991	.932	.962	.957	.961	.992	.890	.959	.953	.949
PTK(c,s_t)	.997	.427	.961	.958	.836	.998	.403	.960	.938	.825
PTK(r,c)+PTK(c,s_t)	.997	.531	.969	.960	.864	.997	.492	.967	.946	.851
SK(r,c)+ PTK(r,c)	.990	.944	.961	.970	.965	.990	.876	.960	.944	.943
SK(r,c)+SK(c,s_t)+ PTK(r,c)+PTK(c,s_t)	.999	.515	.961	.964	.860	.999	.466	.960	.930	.839
mosesBLEU	-	-	-	-	-	.990	.752	.950	.765	.864
mosesWER	-	-	-	-	-	.982	.770	.958	.680	.848
mosesPER	-	-	-	-	-	.981	.770	.974	.947	.918
mosesCDER	-	-	-	-	-	.995	.846	.968	.836	.911
mtevalBLEU	-	-	-	-	-	.992	.858	.962	.899	.928
mtevalNIST	-	-	-	-	-	.988	.924	.966	.952	.958

Table 1: Evaluation results in terms of Pearson correlation for the different TSKM variants. The highlighted TSKM variant indicates the submission to the WMT17 Metrics Task.

sis models [3]. The dependency parse trees obtained were converted to tree representations which can be used by the PTK. The lexical-centered-tree approach presented in Croce et al. (2011) was utilized, which required storing both the grammatical relation and the pos-tag information as the rightmost children of a dependency tree node. The score of the kernel functions were normalized using the formula from Croce et al. (2011):

$$score = \frac{K(T1, T2)}{\sqrt{K(T1, T1) \star K(T2, T2)}} \quad (5)$$

with *T1* and *T2* standing for the input data tuple and *K* indicating the type of kernel function. Regarding SK, only a tokenization of the data was required, as the SK function was applied on substructures composed of the lexical items.

For the computation of the kernel functions we used the Partial Tree Kernel[4] and the Sequence Kernel [5] implementations, found in the KeLP (Kernel-based Learning Platform) (Filice et al., 2015b) (Filice et al., 2015a) library. KeLP is an open source Java platform encompassing kernel based Machine Learning algorithms together with multiple types of kernel functions. The implemented kernels support either vector based input representations or structural ones in the form of trees, sequences or graphs.

[3]https://code.google.com/archive/p/mate-tools/downloads

[4]based on (Moschitti, 2006a)

[5]based on (Bunescu and Mooney, 2005)

4 Evaluation and results

4.1 Experimental setup

The evaluation of TSKM was performed using data pertaining to the News domain from the First Conference On Machine Translation (WMT16) [6]. For the results obtained in the WMT17 Metrics Task, please refer to the official results paper. The following language pairs were used in the evaluation: English-German, Czech-English, German-English, Finnish-English, Russian-English and Turkish-English. The MT outputs evaluated correspond to systems submitted to the WMT16 News Translation Task (Bojar et al., 2016), having different types ranging from statistical phrase-based to neural or syntax-based ones. The test sets consist of approximately 3000 tuples, incorporating the source segment together with the reference and candidate translations. We evaluated TSKM in terms of Pearson correlation with human judgments. During the manual evaluation phase of WTM16, human judgments were collected by ranking five candidate translations, with ties being allowed. In order to compute a single TSKM score for an MT system, all the individual sentence scores were combined by averaging them.

Different variants of TSKM were taken into account for evaluation. To investigate how the lexical variation affects the performance of the metric, we also implemented versions of the metric where lemmas are used instead of the exact lexical items.

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

	not exact		exact	
TSKM	**de-en**	**en-de**	**de-en**	**en-de**
SK(r,c)	.921	.643	.919	.715
SK(c,s_t)	.957	.713	.955	.752
SK(r,c)+SK(c,s_t)	.944	.705	.942	.758
SK(r,c)+SK(s,c_t)	.950	.568	.931	.640
PTK(r,c)	.941	.701	.944	.756
PTK(c,s_t)	.966	.761	.968	.789
PTK(r,c)+PTK(c,s_t)	.957	.750	.960	.792
PTK(r,c)+PTK(s,c_t)	.921	.687	.953	.735
SK(r,c)+ PTK(r,c)	.928	.667	.928	.733
SK(r,c)+SK(c,s_t)+ SK(s,c_t)	.970	.693	.964	.753
PTK(r,c)+PTK(c,s_t)+ PTK(s,c_t)	.979	.770	.973	.810
SK(r,c)+SK(c,s_t)+PTK(r,c)+PTK(c,s_t)	.948	.722	.948	.772
SK(r,c)+SK(s,c_t)+PTK(r,c)+PTK(s,c_t)	.954	.622	.931	.684
SK(r,c)+SK(c,s_t)+SK(s,c_t)+ PTK(r,c)+PTK(c,s_t)+PTK(s,c_t)	.974	.724	.969	.777
mosesBLEU	-	-	.880	.784
mosesWER	-	-	.926	.771
mosesPER	-	-	.843	.681
mosesCDER	-	-	.927	.779
mtevalBLEU	-	-	.905	.752
mtevalNIST	-	-	.887	.625

Table 2: Evaluation results in terms of Pearson correlation for the en-de and de-en language pairs

4.2 Results

The results of the evaluation are presented in Tables 1 and 2, which contain the correlation scores for the different TSKM variants taken into account. For comparison purposes, the scores for some state-of-the-art MT evaluation methods are also presented: BLEU (Papineni et al., 2002), NIST (Doddington, 2002), PER (Tillmann et al., 1997), CDER (Leusch et al., 2006) and WER. The results were obtained using the evaluation scripts made available by the WMT16 conference [7]. The following metric notation was adopted for each of the TSKM variants evaluated: *Kernel*[*level*], where *Kernel* identifies the type of kernel utilized (SK or PTK) and *level* refers to the input data tuple used in the calculation. The possible tuple types are:

- (r,c) - the pair of **r**eference and **c**andidate translations
- (c,s_t) - the pair of **c**andidate translations and translated **s**ource
- (s,c_t) - the pair of **s**ource segment and back-translated **c**andidate

In Table 1, the results of TSKM when applied to the Czech-English, Finnish-English, Russian-English and Turkish-English language pairs are presented. We first experimented with applying TSKM on the (r,c) and the (c,s_t) input data pairs. The best performing TSKM variant, SK(r,c)+ PTK(r,c), represents the combination between PTK and SK applied on the reference and candidate translations. Its average correlation score over all language pairs outperforms the state-of-the-art metrics. We can observe that the addition of the pair consisting of the candidate translation and the pseudo-reference generated mixed results. In the case of Finnish-English there was an obvious downgrade in performance, possibly due to the complex morphology of Finnish. Another observation to be pointed out is that the 'not exact' TSKM variants are stronger correlated with the human judgments than their 'exact' counterparts.

In addition to the metric variants presented in Table 1, we further extended the evaluation to the English-German and German-English language pairs by including the source and back-translation tuple in the evaluation, with the results being presented in Table 2. In this case, the best performing method for both language pairs, PTK(r,c)+PTK(c,s_t)+ PTK(s,c_t), makes use of all the three possible input data tuples, succeeding to outperform the state-of-the-art metrics. Yet another aspect worth to point out is that, in the case of English-German, the 'exact' metric variants are the ones that display better correlations. This would suggest that when choosing between 'not exact' or 'exact' variants for TSKM, the direction of the translation (e.g. in/out of English) should be taken into account. Moreover, we can observe that there is a drastic decrease of correlation in the case of English-German translations, which can possibly be explained by the highly inflectional nature of the German language.

	exact	
TSKM	**tr-en**	**ru-en**
SPTK(r,c)	.976	.970
SPTK(c,s_t)	.960	.964
CSPTK(r,c)	.972	.968
CSPTK(c,s_t)	.968	.960

Table 3: Evaluation results in terms of Pearson correlation for SPTK and CSPTK.

Additional preliminary evaluation experiments, presented in Table 3, were performed after the submission to the Shared Task. Generalizations of the Partial Tree Kernel were used, namely the Smoothed Partial Tree Kernel (SPTK) (Croce

[7] http://www.statmt.org/wmt16/results.html

et al., 2011) and the Compositional Smoothed Partial Tree Kernel (CSPTK) (Annesi et al., 2013) (Annesi et al., 2014). The SPTK uses a term similarity function to semantically match tree nodes. The term similarity function can be obtained through either word vector spaces or distributional analysis. On the other hand, the CSPTK represents a generalization of SPTK, which uses Distributional Compositional Semantics to determine the degree of similarity between tree fragments. The implementations for these kernels together with an example wordspace for English are also available in the KeLP package. The results show that by relaxing the matching constraints to allow for lexical variation these kernels outperform PTK when used by TSKM.

5 Conclusions and future work

In this paper, we introduced TSKM, our submission to the WMT17 Metrics Task, which is based on tree and sequence kernels. The metric was evaluated using multiple language pairs, with the evaluation results being very encouraging. We also experimented with applying the kernel functions on additional tuple input data, that involve back-translations and pseudo-references. In the case of the pseudo-reference the results indicate that its addition to TSKM can be beneficial, especially in the case of the PTK. However, the most important aspect to notice is that, with the exception of Finnish-English, the pseudo-reference based methods achieved correlation scores that are very similar to the official reference based ones, which suggests that TSKM could be applied even in the context of artificially generated reference translations. The addition of the back-translations of the target sentences to TSKM generated encouraging results, which prompts us to extend the evaluation to include further language pairs.

Based on the evaluation results, we can also observe that the SK metric variants succeeded in attaining correlation scores that are relatively similar to the PTK variants. This suggests that the SK metric variant can be successfully used in the case when no syntactic analysis tools are available for the target language.

Future work will be concentrated on using the constituency trees as a structural input representations for PTK in addition to the dependency trees. The evaluation will also be extended to determine how well does TSKM generalize across domains. We also plan to analyze in more detail the decrease in correlation scores when using the pseudo-reference in the case of Finnish-English, by using different MT systems to generate additional pseudo-references in order to determine if the type of MT system influences the correlation with human judgments. Another future work idea is to extend the evaluation for SPTK and CSPTK, by including them in different TSKM combinations and evaluating on additional language pairs.

References

Joshua Albrecht and Rebecca Hwa. 2007. Regression for sentence-level MT evaluation with pseudo references. In *Annual Meeting-Association for Computational Linguistics*. volume 45, page 296.

Joshua Albrecht and Rebecca Hwa. 2008. The role of pseudo references in MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 187–190.

Waleed Ammar, Victor Chahuneau, Michael Denkowski, Greg Hanneman, Wang Ling, Austin Matthews, Kenton Murray, Nicola Segall, Yulia Tsvetkov, Alon Lavie, and Chris Dyer. 2013. The CMU Machine Translation systems at WMT 2013: Syntax, synthetic translation options, and pseudo-references. In *8th Workshop on Statistical Machine Translation*. page 70.

Paolo Annesi, Danilo Croce, and Roberto Basili. 2013. Towards compositional tree kernels. In *Joint Symposium on Semantic Processing.*. page 15.

Paolo Annesi, Danilo Croce, and Roberto Basili. 2014. Semantic compositionality in tree kernels. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, pages 1029–1038.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. *The 23rd International Conference on Computational Linguistics (COLING 2010)* .

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers* pages 131–198.

Razvan Bunescu and Raymond Mooney. 2005. Subsequence kernels for Relation Extraction. *Advances*

in Neural Information Processing Systems, Vol. 18: Proceedings of the 2005 Conference (NIPS) .

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in Machine Translation research. In *In EACL*. pages 249–256.

Julio Castillo and Paula Estrella. 2012. Semantic Textual Similarity for MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 52–58.

Michael Collins and Nigel Duffy. 2001. Convolution kernels for natural language. *Proceedings of NIPS 2001* pages 625–632.

Michael Collins and Nigel Duffy. 2002. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 263–270.

Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2011. Structured lexical similarity via convolution kernels on dependency trees. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* pages 1034–1046.

George Doddington. 2002. Automatic evaluation of Machine Translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., pages 138–145.

Simone Filice, Giuseppe Castellucci, Roberto Basili, Giovanni Da San Martino, and Alessandro Moschitti. 2015a. KeLP: a Kernel-based Learning Platform in Java. *The workshop on Machine Learning Open Source Software (MLOSS): Open Ecosystems* .

Simone Filice, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2015b. KeLP: a kernel-based learning platform for Natural Language Processing. *Proceedings of ACL-IJCNLP 2015 System Demonstrations* pages 19–24.

Shubham Gautam and Pushpak Bhattacharyya. 2014. Layered: Metric for Machine Translation evaluation. *ACL 2014* page 387.

Kim Gerdes. 2013. Collaborative dependency annotation. In *DepLing*. pages 88–97.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic features for automatic evaluation of heterogenous MT systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 256–264.

Francisco Guzmán, Shafiq R Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves Machine Translation evaluation. In *ACL (1)*. pages 687–698.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. Efficient MT evaluation using block movements. In *Proceedings of EACL-2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*. pages 241–248.

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.

Alessandro Moschitti. 2006a. Efficient convolution kernels for dependency and constituent syntactic trees. *Proceedings of the 17th European Conference on Machine Learning* .

Truc-Vien T. Nguyen, Alessandro Moschitti, and Giuseppe Riccardi. 2009. Convolution kernels on constituent, dependency and sequential structures for Relation Extraction. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* pages 1378–1387.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Maja Popović and Hermann Ney. 2007. Word Error Rates: Decomposition over POS classes and applications for error analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 48–55.

Maja Popović and Hermann Ney. 2009. Syntax-oriented evaluation measures for Machine Translation output. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 29–32.

Reinhard Rapp. 2009. The back-translation score: Automatic MT evaluation at the sentence level without reference translations. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, pages 133–136.

M Snover, BJ Dorr, R Schwartz, L Micciulla, and J Makhoul. 2006. A study of translation edit rate with targeted human annotation. 2006. In *Proc. AMTA*.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In *Eurospeech*.