DiscoTK: Using Discourse Structure for Machine Translation Evaluation

Shafiq Joty **Francisco Guzmán** **Lluís Màrquez** and **Preslav Nakov**
ALT Research Group
Qatar Computing Research Institute — Qatar Foundation
{sjoty,fguzman,lmarquez,pnakov}@qf.org.qa

Abstract

We present novel automatic metrics for machine translation evaluation that use discourse structure and convolution kernels to compare the discourse tree of an automatic translation with that of the human reference. We experiment with five transformations and augmentations of a base discourse tree representation based on the rhetorical structure theory, and we combine the kernel scores for each of them into a single score. Finally, we add other metrics from the ASIYA MT evaluation toolkit, and we tune the weights of the combination on actual human judgments. Experiments on the WMT12 and WMT13 metrics shared task datasets show correlation with human judgments that outperforms what the best systems that participated in these years achieved, both at the segment and at the system level.

1 Introduction

The rapid development of statistical machine translation (SMT) that we have seen in recent years would not have been possible without automatic metrics for measuring SMT quality. In particular, the development of BLEU (Papineni et al., 2002) revolutionized the SMT field, allowing not only to compare two systems in a way that strongly correlates with human judgments, but it also enabled the rise of discriminative log-linear models, which use optimizers such as MERT (Och, 2003), and later MIRA (Watanabe et al., 2007; Chiang et al., 2008) and PRO (Hopkins and May, 2011), to optimize BLEU, or an approximation thereof, directly. While over the years other strong metrics such as TER (Snover et al., 2006) and Meteor (Lavie and Denkowski, 2009) have emerged, BLEU remains the de-facto standard, despite its simplicity.

Recently, there has been steady increase in BLEU scores for well-resourced language pairs such as Spanish-English and Arabic-English. However, it was also observed that BLEU-like n-gram matching metrics are unreliable for high-quality translation output (Doddington, 2002; Lavie and Agarwal, 2007). In fact, researchers already worry that BLEU will soon be unable to distinguish automatic from human translations.[1] This is a problem for most present-day metrics, which cannot tell apart raw machine translation output from a fully fluent professionally post-edited version thereof (Denkowski and Lavie, 2012).

Another concern is that BLEU-like n-gram matching metrics tend to favor phrase-based SMT systems over rule-based systems and other SMT paradigms. In particular, they are unable to capture the syntactic and semantic structure of sentences, and are thus insensitive to improvement in these aspects. Furthermore, it has been shown that lexical similarity is both insufficient and not strictly necessary for two sentences to convey the same meaning (Culy and Riehemann, 2003; Coughlin, 2003; Callison-Burch et al., 2006).

The above issues have motivated a large amount of work dedicated to design better evaluation metrics. The Metrics task at the Workshop on Machine Translation (WMT) has been instrumental in this quest. Below we present QCRI's submission to the Metrics task of WMT14, which consists of the DiscoTK family of discourse-based metrics.

In particular, we experiment with five different transformations and augmentations of a discourse tree representation, and we combine the kernel scores for each of them into a single score which we call DISCOTK$_{light}$. Next, we add to the combination other metrics from the ASIYA MT evaluation toolkit (Giménez and Màrquez, 2010), to produce the DISCOTK$_{party}$ metric.

[1] This would not mean that computers have achieved human proficiency; it would rather show BLEU's inadequacy.

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 402–408,
Baltimore, Maryland USA, June 26–27, 2014.

Finally, we tune the relative weights of the metrics in the combination using human judgments in a learning-to-rank framework. This proved to be quite beneficial: the tuned version of the DISCOTK$_{party}$ metric was the best performing metric in the WMT14 Metrics shared task.

The rest of the paper is organized as follows: Section 2 introduces our basic discourse metrics and the tree representations they are based on. Section 3 describes our metric combinations. Section 4 presents our experiments and results on datasets from previous years. Finally, Section 5 concludes and suggests directions for future work.

2 Discourse-Based Metrics

In our recent work (Guzmán et al., 2014), we used the information embedded in the discourse-trees (DTs) to compare the output of an MT system to a human reference. More specifically, we used a state-of-the-art sentence-level discourse parser (Joty et al., 2012) to generate discourse trees for the sentences in accordance with the Rhetorical Structure Theory (RST) of discourse (Mann and Thompson, 1988). Then, we computed the similarity between DTs of the human references and the system translations using a convolution tree kernel (Collins and Duffy, 2001), which efficiently computes the number of common subtrees. Note that this kernel was originally designed for syntactic parsing, and the subtrees are subject to the constraint that their nodes are taken with all or none of their children, i.e., if we take a direct descendant of a given node, we must also take all siblings of that descendant. This imposes some limitations on the type of substructures that can be compared, and motivates the enriched tree representations explained in subsections 2.1–2.4.

The motivation to compare discourse trees, is that translations should preserve the coherence relations. For example, consider the three discourse trees (DTs) shown in Figure 1. Notice that the *Attribution* relation in the reference translation is also realized in the system translation in (b) but not in (c), which makes (b) a better translation compared to (c), according to our hypothesis.

In (Guzmán et al., 2014), we have shown that discourse structure provides additional information for MT evaluation, which is not captured by existing metrics that use lexical, syntactic and semantic information; thus, discourse should be considered when developing new rich metrics.

Here, we extend our previous work by developing metrics that are based on new representations of the DTs. In the remainder of this section, we will focus on the individual DT representations that we will experiment with; then, the following section will describe the metric combinations and tuning used to produce the DiscoTK metrics.

2.1 DR-LEX$_1$

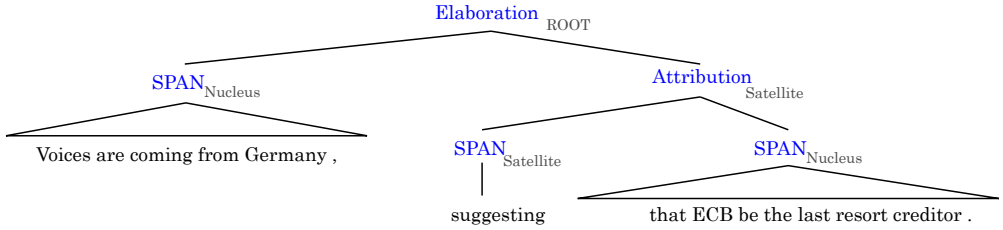
Figure 2a shows our first representation of the DT. The lexical items, i.e., words, constitute the leaves of the tree. The words in an Elementary Discourse Unit (EDU) are grouped under a predefined tag **EDU**, to which the nuclearity status of the EDU is attached: *nucleus* vs. *satellite*. Coherence relations, such as *Attribution*, *Elaboration*, and *Enablement*, between adjacent text spans constitute the internal nodes of the tree. Like the EDUs, the nuclearity statuses of the larger discourse units are attached to the relation labels. Notice that with this representation the tree kernel can easily be extended to find subtree matches at the word level, i.e., by including an additional layer of *dummy* leaves as was done in (Moschitti et al., 2007). We applied the same solution in our representations.

2.2 DR-NOLEX

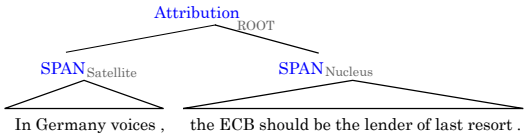
Our second representation DR-NOLEX (Figure 2b) is a simple variation of DR-LEX$_1$, where we exclude the lexical items. This allows us to measure the similarity between two translations in terms of their discourse structures alone.

2.3 DR-LEX$_2$

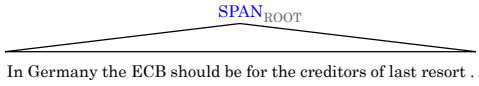
One limitation of DR-LEX$_1$ and DR-NOLEX is that they do not separate the structure, i.e., the skeleton, of the tree from its labels. Therefore, when measuring the similarity between two DTs, they do not allow the tree kernel to give partial credit to subtrees that differ in labels but match in their structures. DR-LEX$_2$, a variation of DR-LEX$_1$, addresses this limitation as shown in Figure 2c. It uses predefined tags **SPAN** and **EDU** to build the skeleton of the tree, and considers the nuclearity and/or relation labels as properties (added as children) of these tags. For example, a **SPAN** has two properties, namely its nuclearity and its relation, and an **EDU** has one property, namely its nuclearity. The words of an EDU are placed under the predefined tag **NGRAM**.



(a) A reference (human-written) translation.



(b) A higher quality (system-generated) translation.



(c) A lower quality (system-generated) translation.

Figure 1: Three discourse trees for the translations of a source sentence: (a) the reference, (b) a higher quality automatic translation, and (c) a lower quality automatic translation.

2.4 DR-LEX$_{1.1}$ and DR-LEX$_{2.1}$

Although both DR-LEX$_1$ and DR-LEX$_2$ allow the tree kernel to find matches at the word level, the words are compared in a bag-of-words fashion, i.e., if the trees share a common word, the kernel will find a match regardless of its position in the tree. Therefore, a word that has occurred in an EDU with status *Nucleus* in one tree could be matched with the same word under a *Satellite* in the other tree. In other words, the kernel based on these representations is insensitive to the nuclearity status and the relation labels under which the words are matched. DR-LEX$_{1.1}$, an extension of DR-LEX$_1$, and DR-LEX$_{2.1}$, an extension of DR-LEX$_2$, are sensitive to these variations at the lexical level. DR-LEX$_{1.1}$ (Figure 2d) and DR-LEX$_{2.1}$ (Figure 2e) propagate the nuclearity statuses and/or the relation labels to the lexical items by including three more subtrees at the EDU level.

3 Metric Combination and Tuning

In this section, we describe our Discourse Tree Kernel (DiscoTK) metrics. We have two main versions: DISCOTK$_{light}$, which combines the five DR-based metrics, and DISCOTK$_{party}$, which further adds the Asiya metrics.

3.1 DISCOTK$_{light}$

In the previous section, we have presented several discourse tree representations that can be used to compare the output of a machine translation system to a human reference. Each representation stresses a different aspect of the discourse tree.

In order to make our estimations more robust, we propose DISCOTK$_{light}$, a metric that takes advantage of all the previous discourse representations by linearly interpolating their scores. Here are the processing steps needed to compute this metric:

(*i*) Parsing: We parsed each sentence in order to produce discourse trees for the human references and for the outputs of the systems.

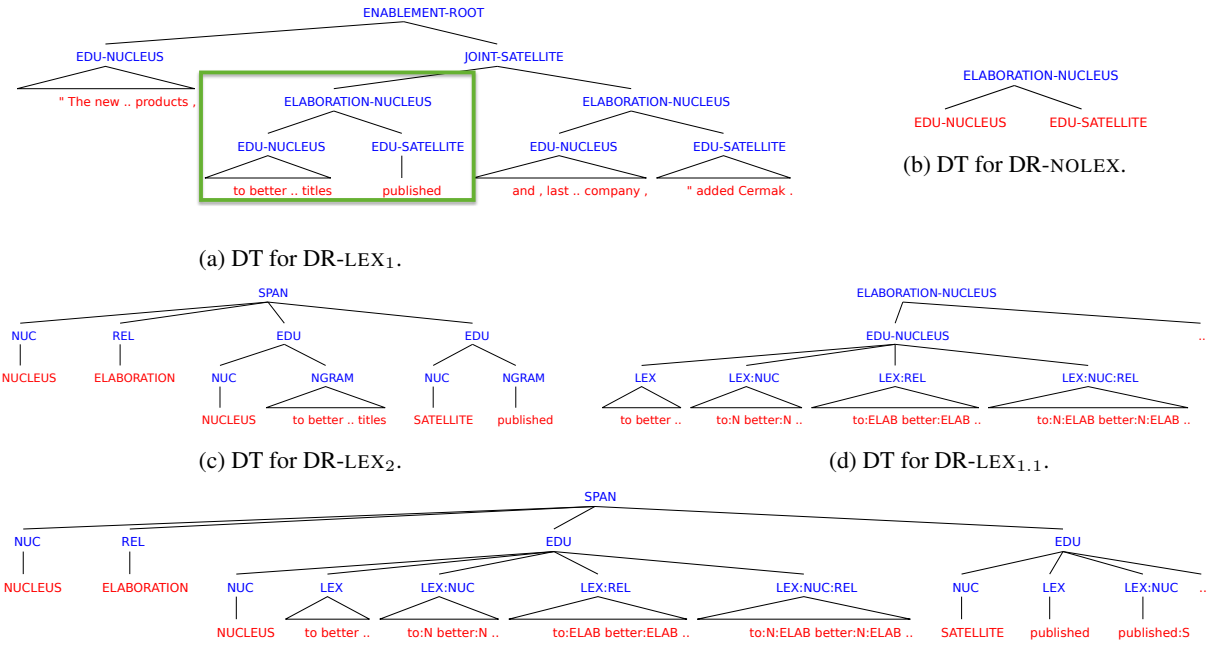
(*ii*) Tree enrichment/simplification: For each sentence-level discourse tree, we generated the five different tree representations: DR-NOLEX, DR-LEX$_1$, DR-LEX$_{1.1}$, DR-LEX$_2$, DR-LEX$_{2.1}$.

(*iii*) Estimation: We calculated the per-sentence similarity scores between tree representations of the system hypothesis and the human reference using the extended convolution tree kernel as described in the previous section. To compute the system-level similarity scores, we calculated the average sentence-level similarity; note that this ensures that our metric is "the same" at the system and at the segment level.

(*iv*) Normalization: In order to make the scores of the different representations comparable, we performed a min–max normalization[2] for each metric and for each language pair.

(*v*) Combination: Finally, for each sentence, we computed DISCOTK$_{light}$ as the average of the normalized similarity scores of the different representations. For system-level experiments, we performed linear interpolation of system-level scores.

[2]Where $x' = (x - \min)/(\max - \min)$.



(a) DT for DR-LEX$_1$.

(b) DT for DR-NOLEX.

(c) DT for DR-LEX$_2$.

(d) DT for DR-LEX$_{1.1}$.

(e) DT for DR-LEX$_{2.1}$.

Figure 2: Five different representations of the discourse tree (DT) for the sentence *"The new organisational structure will also allow us to enter the market with a joint offer of advertising products, to better link the creation of content for all the titles published and, last but not least, to continue to streamline significantly the business management of the company," added Cermak.* Note that to avoid visual clutter, (b)–(e) show alternative representations only for the highlighted subtree in (a).

3.2 DISCOTK$_{party}$

One of the weaknesses of the above discourse-based metrics is that they use unigram lexical information, which does not capture reordering. Thus, in order to make more informed and robust estimations, we extended DISCOTK$_{light}$ with the composing metrics of the ASIYA's ULC metric (Giménez and Màrquez, 2010), which is a uniform linear combination of twelve individual metrics and was the best-performing metric at the system and at the segment levels at the WMT08 and WMT09 metrics tasks.

In order to compute the individual metrics from ULC, we used the ASIYA toolkit,[3] but we departed from ASIYA's ULC by replacing TER and Meteor with newer versions thereof that take into account synonymy lookup and paraphrasing ('TERp-A' and 'Meteor-pa' in ASIYA's terminology). We then combined the five components in DISCOTK$_{light}$ and the twelve individual metrics from ULC; we call this combination DISCOTK$_{party}$.

[3]http://nlp.lsi.upc.edu/asiya/

We combined the scores using linear interpolation in two different ways:

(*i*) *Uniform combination* of min-max normalized scores at the segment level. We obtained system-level scores by computing the average over the segment scores.

(*ii*) *Trained interpolation at the sentence level.* We determined the interpolation weights for the above-described combination of 5+12 = 17 metrics using a pairwise learning-to-rank framework and classification with logistic regression, as we had done in (Guzmán et al., 2014). We obtained the final test-time sentence-level scores by passing the interpolated raw scores through a sigmoid function. In contrast, for the final system-level scores, we averaged the per-sentence interpolated raw scores.

We also tried to learn the interpolation weights at the system level, experimenting with both regression and classification. However, the amount of data available for this type of training was small, and the learned weights did not perform significantly better than the uniform combination.

3.3 Post-processing

Discourse-based metrics, especially DR-NOLEX, tend to produce many ties when there is not enough information to do complete discourse analysis. This contributes to lower τ scores for DISCOTK$_{light}$. To alleviate this issue, we used a simple tie-breaking strategy, in which ties between segment scores for different systems are resolved by using perturbations proportional to the global system-level scores produced by the same metric, i.e., $x'^{seg}_{sys} = x^{seg}_{sys} + \epsilon * \sum_s x^s_{sys}$. Here, ϵ is automatically chosen to avoid collisions with scores not involved in the tie. This post-processing is not part of the metric; it is only applied to our segment-level submission to the WMT'14 metrics task.

4 Experimental Evaluation

In this section, we present some of our experiments to decide on the best DiscoTK metric variant and tuning set. For tuning, testing and comparison, we worked with some of the datasets available from previous WMT metrics shared tasks, i.e., 2011, 2012 and 2013. From previous experiments (Guzmán et al., 2014), we know that the tuned metrics perform very well on cross-validation for the same-year dataset. We further know that tuning can be performed by concatenating data from all the into-English language pairs, which yields better results than training separately by language pair. For the WMT14 metrics task, we investigated in more depth whether the tuned metrics generalize well to new datasets. Additionally, we tested the effect of concatenating datasets from different years.

Table 1 shows the main results of our experiments with the DiscoTK metrics. We evaluated the performance of the metrics on the WMT12 and WMT13 datasets both at the segment and the system level, and we used WMT11 as an additional tuning dataset. We measured the performance of the metrics in terms of correlation with human judgements. At the segment level, we evaluated using Kendall's Tau (τ), recalculated following the WMT14 official Kendall's Tau implementation. At the system level, we used Spearman's rank correlation (ρ) and Pearson's correlation coefficient (r). In all cases, we averaged the results over all into-English language pairs. The symbol '$\varnothing$' represents the untuned versions of our metrics, i.e., applying a uniform linear combination of the individual metrics.

We trained the tuned versions of the DiscoTK measures using different datasets (WMT11, WMT12 and WMT13) in order to study across-corpora generalization and the effect of training dataset size. The symbol '+' stands for concatenation of datasets. We trained the tuned versions at the segment level using Maximum Entropy classifiers for pairwise ranking (cf. Section 3). For the sake of comparison, the first group of rows contains the results of the best-performing metrics at the WMT12 and WMT13 metrics shared tasks and the last group of rows contains the results of the ASIYA combination of metrics, i.e., DISCOTK$_{party}$ without the discourse components.

Several conclusions can be drawn from Table 1. First, DISCOTK$_{party}$ is better than DISCOTK$_{light}$ in all settings, indicating that the discourse-based metrics are very well complemented by the heterogeneous metric set from ASIYA. DISCOTK$_{light}$ achieves competitive scores at the system level (which would put the metric among the best participants in WMT12 and WMT13); however, as expected, it is not robust enough at the segment level. On the other hand, the tuned versions of DISCOTK$_{party}$ are very competitive and improve over the already strong ASIYA in each configuration both at the segment- and the system-level. The improvements are small but consistent, showing that using discourse increases the correlation with human judgments.

Focusing on the results at the segment level, it is clear that the tuned versions offer an advantage over the simple uniform linear combinations. Interestingly, for the tuned variants, given a test set, the results are consistent across tuning sets, ruling out over-fitting; this shows that the generalization is very good. This result aligns well with what we observed in our previous studies (Guzmán et al., 2014). Learning with more data (WMT11+12 or WMT12+13) does not seem to help much, but it does not hurt performance either. Overall, the τ correlation results obtained with the tuned DISCOTK$_{party}$ metric are much better than the best results of any participant metrics at WMT12 and WMT13 (20.1% and 9.5% relative improvement, respectively).

At the system level, we observe that tuning over the DISCOTK$_{light}$ metric is not helpful (results are actually slightly lower), while tuning the more complex DISCOTK$_{party}$ metric yields slightly better results.

		Segment Level		System Level			
		WMT12	WMT13	WMT12		WMT13	
Metric	Tuning	τ	τ	ρ	r	ρ	r
SEMPOS	na	–	–	0.902	0.922	–	–
SPEDE07PP	na	0.254	–	–	–	–	–
METEOR-WMT13	na	–	0.264	–	–	0.935	**0.950**
DISCOTK$_{light}$	∅	0.171	0.162	0.884	0.922	0.880	0.911
	WMT11	0.207	0.201	0.860	0.872	0.890	0.909
	WMT12	–	0.200	–	–	0.889	0.910
	WMT13	0.206	–	0.865	0.871	–	–
	WMT11+12	–	0.197	–	–	0.890	0.910
	WMT11+13	0.207	–	0.865	0.871	–	–
DISCOTK$_{party}$	∅	0.257	0.231	0.907	0.915	**0.941**	0.928
	WMT11	0.302	0.282	**0.915**	**0.940**	0.934	0.946
	WMT12	–	0.284	–	–	0.936	0.940
	WMT13	**0.305**	–	0.912	0.935	–	–
	WMT11+12	–	**0.289**	–	–	0.936	0.943
	WMT11+13	0.304	–	0.912	0.934	–	–
ASIYA	∅	0.273	0.252	0.899	0.909	0.932	0.922
	WMT11	0.301	0.279	0.913	0.935	0.934	0.944
	WMT12	–	0.277	–	–	0.932	0.938
	WMT13	0.303	–	0.908	0.932	–	–
	WMT11+12	–	0.277	–	–	0.934	0.940
	WMT11+13	0.303	–	0.908	0.933	–	–

Table 1: Evaluation results on WMT12 and WMT13 datasets at segment and system level for the main combined DiscoTK measures proposed in this paper.

The scores of our best metric are higher than those of the best participants in WMT12 and WMT13, according to Spearman's ρ, which was the official metric in those years. Overall, our metrics are comparable to the state-of-the-art at the system level. The differences between Spearman's ρ and Pearson's r coefficients are not dramatic, with r values being always higher than ρ.

Given the above results, we submitted the following runs to the WMT14 Metrics shared task: (*i*) DISCOTK$_{party}$ tuned on the concatenation of datasets WMT11+12+13, as our primary run; (*ii*) Untuned DISCOTK$_{party}$, to verify that we are not over-fitting the training set; and (*iii*) Untuned DISCOTK$_{light}$, to see the performance of a metric using discourse structures and word unigrams.

The results for the WMT14 Metrics shared task have shown that our primary run, DISCOTK$_{party}$ tuned, was the *best-performing* metric both at the segment- and at the system-level (Macháček and Bojar, 2014). This metric yielded significantly better results than its untuned counterpart, confirming the importance of weight tuning and the absence of over-fitting during tuning. Finally, the untuned DISCOTK$_{light}$ achieved relatively competitive, albeit slightly worse results for all language pairs, except for Hindi-English, where system translations resembled a "word salad", and were very hard to discourse-parse accurately.

5 Conclusion

We have presented experiments with novel automatic metrics for machine translation evaluation that take discourse structure into account. In particular, we used RST-style discourse parse trees, which we compared using convolution kernels. We further combined these kernels with metrics from ASIYA, also tuning the weights. The resulting DISCOTK$_{party}$ tuned metric was the best-performing at the segment- and system-level at the WMT14 metrics task.

In an internal evaluation on the WMT12 and WMT13 metrics datasets, this tuned combination showed correlation with human judgments that outperforms the best systems that participated in these shared tasks. The discourse-only metric ranked near the top at the system-level for WMT12 and WMT13; however, it is weak at the segment-level since it is sensitive to parsing errors, and most sentences have very little internal discourse structure.

In the future, we plan to work on an integrated representation of syntactic, semantic and discourse-based tree structures, which would allow us to design evaluation metrics based on more fine-grained features, and would also allow us to train such metrics using kernel methods. Furthermore, we want to make use of discourse parse information beyond the sentence level.

References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*, EACL'06, pages 249–256, Trento, Italy.

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'08, pages 224–233, Honolulu, Hawaii.

Michael Collins and Nigel Duffy. 2001. Convolution Kernels for Natural Language. In *Neural Information Processing Systems*, NIPS'01, pages 625–632, Vancouver, Canada.

Deborah Coughlin. 2003. Correlating automated and human assessments of machine translation quality. In *Proceedings of the Machine Translation Summit IX*, MT Summit'03, pages 23–27, New Orleans, LA, USA.

Christopher Culy and Susanne Riehemann. 2003. The limits of n-gram translation evaluation metrics. In *Proceedings of the Machine Translation Summit IX*, MT Summit'03, pages 1–8, New Orleans, LA, USA.

Michael Denkowski and Alon Lavie. 2012. Challenges in predicting machine translation utility for human post-editors. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*, AMTA'12, pages 40–49, San Diego, CA, USA.

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*, HLT'02, pages 138–145, San Francisco, CA, USA.

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

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, ACL'14, Baltimore, MD, USA.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'11, pages 1352–1362, Edinburgh, Scotland, UK.

Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2012. A Novel Discriminative Framework for Sentence-Level Discourse Analysis. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL'12, pages 904–915, Jeju Island, Korea.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, WMT'07, pages 228–231, Prague, Czech Republic.

Alon Lavie and Michael Denkowski. 2009. The METEOR metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 Metrics Shared Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, WMT'14, Baltimore, MD, USA.

William Mann and Sandra Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281.

Alessandro Moschitti, Silvia Quarteroni, Roberto Basili, and Suresh Manandhar. 2007. Exploiting syntactic and shallow semantic kernels for question answer classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL'07, pages 776–783, Prague, Czech Republic.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, ACL'03, pages 160–167, Sapporo, Japan.

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*, ACL'02, pages 311–318, Philadelphia, PA, USA.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Seventh Biennial Conference of the Association for Machine Translation in the Americas*, AMTA'06, pages 223–231, Cambridge, MA, USA.

Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki. 2007. Online large-margin training for statistical machine translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL'07, pages 764–773, Prague, Czech Republic.