Using Discourse Structure Improves Machine Translation Evaluation

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

Abstract

We present experiments in using discourse structure for improving machine translation evaluation. We first design two discourse-aware similarity measures, which use all-subtree kernels to compare discourse parse trees in accordance with the Rhetorical Structure Theory. Then, we show that these measures can help improve a number of existing machine translation evaluation metrics both at the segment- and at the system-level. Rather than proposing a single new metric, we show that discourse information is complementary to the state-of-the-art evaluation metrics, and thus should be taken into account in the development of future richer evaluation metrics.

1 Introduction

From its foundations, Statistical Machine Translation (SMT) had two defining characteristics: first, translation was modeled as a generative process *at the sentence-level*. Second, it was purely statistical over words or word sequences and made *little to no use of linguistic information*. Although modern SMT systems have switched to a discriminative log-linear framework, which allows for additional sources as features, it is generally hard to incorporate dependencies beyond a small window of adjacent words, thus making it difficult to use linguistically-rich models.

Recently, there have been two promising research directions for improving SMT and its evaluation: (*a*) by using more structured linguistic information, such as syntax (Galley et al., 2004; Quirk et al., 2005), hierarchical structures (Chiang, 2005), and semantic roles (Wu and Fung, 2009; Lo et al., 2012), and (*b*) by going beyond the sentence-level, e.g., translating at the document level (Hardmeier et al., 2012).

Going beyond the sentence-level is important since sentences rarely stand on their own in a well-written text. Rather, each sentence follows smoothly from the ones before it, and leads into the ones that come afterwards. The logical relationship between sentences carries important information that allows the text to express a meaning as a whole beyond the sum of its separate parts.

Note that sentences can be made of several clauses, which in turn can be interrelated through the same logical relations. Thus, in a coherent text, discourse units (sentences or clauses) are logically connected: the meaning of a unit relates to that of the previous and the following units.

Discourse analysis seeks to uncover this coherence structure underneath the text. Several formal theories of discourse have been proposed to describe the coherence structure (Mann and Thompson, 1988; Asher and Lascarides, 2003; Webber, 2004). For example, the Rhetorical Structure Theory (Mann and Thompson, 1988), or RST, represents text by labeled hierarchical structures called Discourse Trees (DTs), which can incorporate several layers of other linguistic information, e.g., syntax, predicate-argument structure, etc.

Modeling discourse brings together the above research directions (a) and (b), which makes it an attractive goal for MT. This is demonstrated by the establishment of a recent workshop dedicated to Discourse in Machine Translation (Webber et al., 2013), collocated with the 2013 annual meeting of the Association of Computational Linguistics.

The area of discourse analysis for SMT is still nascent and, to the best of our knowledge, no previous research has attempted to use rhetorical structure for SMT or machine translation evaluation. One possible reason could be the unavailability of accurate discourse parsers. However, this situation is likely to change given the most recent advances in automatic discourse analysis (Joty et al., 2012; Joty et al., 2013).

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 687–698,
Baltimore, Maryland, USA, June 23-25 2014.

We believe that the semantic and pragmatic information captured in the form of DTs (*i*) can help develop discourse-aware SMT systems that produce coherent translations, and (*ii*) can yield better MT evaluation metrics. While in this work we focus on the latter, we think that the former is also within reach, and that SMT systems would benefit from preserving the coherence relations in the source language when generating target-language translations.

In this paper, rather than proposing yet another MT evaluation metric, we show that discourse information is complementary to many existing evaluation metrics, and thus should not be ignored. We first design two discourse-aware similarity measures, which use DTs generated by a publicly-available discourse parser (Joty et al., 2012); then, we show that they can help improve a number of MT evaluation metrics at the segment- and at the system-level in the context of the WMT11 and the WMT12 metrics shared tasks (Callison-Burch et al., 2011; Callison-Burch et al., 2012).

These metrics tasks are based on sentence-level evaluation, which arguably can limit the benefits of using global discourse properties. Fortunately, several sentences are long and complex enough to present rich discourse structures connecting their basic clauses. Thus, although limited, this setting is able to demonstrate the potential of discourse-level information for MT evaluation. Furthermore, sentence-level scoring (*i*) is compatible with most translation systems, which work on a sentence-by-sentence basis, (*ii*) could be beneficial to modern MT tuning mechanisms such as PRO (Hopkins and May, 2011) and MIRA (Watanabe et al., 2007; Chiang et al., 2008), which also work at the sentence-level, and (*iii*) could be used for re-ranking n-best lists of translation hypotheses.

2 Related Work

Addressing discourse-level phenomena in machine translation is relatively new as a research direction. Some recent work has looked at anaphora resolution (Hardmeier and Federico, 2010) and discourse connectives (Cartoni et al., 2011; Meyer, 2011), to mention two examples.[1] However, so far the attempts to incorporate discourse-related knowledge in MT have been only moderately successful, at best.

[1]We refer the reader to (Hardmeier, 2012) for an in-depth overview of discourse-related research for MT.

A common argument, is that current automatic evaluation metrics such as BLEU are inadequate to capture discourse-related aspects of translation quality (Hardmeier and Federico, 2010; Meyer et al., 2012). Thus, there is consensus that discourse-informed MT evaluation metrics are needed in order to advance research in this direction. Here we suggest some simple ways to create such metrics, and we also show that they yield better correlation with human judgments.

The field of automatic evaluation metrics for MT is very active, and new metrics are continuously being proposed, especially in the context of the evaluation campaigns that run as part of the Workshops on Statistical Machine Translation (WMT 2008-2012), and NIST Metrics for Machine Translation Challenge (MetricsMATR), among others. For example, at WMT12, 12 metrics were compared (Callison-Burch et al., 2012), most of them new.

There have been several attempts to incorporate syntactic and semantic linguistic knowledge into MT evaluation. For instance, at the syntactic level, we find metrics that measure the structural similarity between shallow syntactic sequences (Giménez and Màrquez, 2007; Popovic and Ney, 2007) or between constituency trees (Liu and Gildea, 2005). In the semantic case, there are metrics that exploit the similarity over named entities and predicate-argument structures (Giménez and Màrquez, 2007; Lo et al., 2012).

In this work, instead of proposing a new metric, we focus on enriching current MT evaluation metrics with discourse information. Our experiments show that many existing metrics can benefit from additional knowledge about discourse structure.

In comparison to the syntactic and semantic extensions of MT metrics, there have been very few attempts to incorporate discourse information so far. One example are the semantics-aware metrics of Giménez and Màrquez (2009) and Comelles et al. (2010), which use the Discourse Representation Theory (Kamp and Reyle, 1993) and tree-based discourse representation structures (DRS) produced by a semantic parser. They calculate the similarity between the MT output and references based on DRS subtree matching, as defined in (Liu and Gildea, 2005), DRS lexical overlap, and DRS morpho-syntactic overlap. However, they could not improve correlation with human judgments, as evaluated on the MetricsMATR dataset.

Compared to the previous work, (*i*) we use a different discourse representation (RST), (*ii*) we compare discourse parses using *all-subtree* kernels (Collins and Duffy, 2001), (*iii*) we evaluate on much larger datasets, for several language pairs and for multiple metrics, and (*iv*) we do demonstrate better correlation with human judgments.

Wong and Kit (2012) recently proposed an extension of MT metrics with a measure of document-level *lexical cohesion* (Halliday and Hasan, 1976). Lexical cohesion is achieved using word repetitions and semantically similar words such as synonyms, hypernyms, and hyponyms. For BLEU and TER, they observed improved correlation with human judgments on the MTC4 dataset when linearly interpolating these metrics with their *lexical cohesion* score. Unlike their work, which measures lexical cohesion at the document-level, here we are concerned with *coherence (rhetorical) structure*, primarily at the sentence-level.

3 Our Discourse-Based Measures

Our working hypothesis is that the similarity between the discourse structures of an automatic and of a reference translation provides additional information that can be valuable for evaluating MT systems. In particular, we believe that good translations should tend to preserve discourse relations.

As an example, consider the three discourse trees (DTs) shown in Figure 1: (*a*) for a reference (human) translation, and (*b*) and (*c*) for translations of two different systems on the WMT12 test dataset. The leaves of a DT correspond to contiguous atomic text spans, called *Elementary Discourse Units* or EDUs (three in Figure 1a). Adjacent spans are connected by certain coherence relations (e.g., *Elaboration*, *Attribution*), forming larger discourse units, which in turn are also subject to this relation linking. Discourse units linked by a relation are further distinguished based on their relative importance in the text: *nuclei* are the core parts of the relation while *satellites* are supportive ones. Note that the nuclearity and relation labels in the reference translation are 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. We argue that existing metrics that only use lexical and syntactic information cannot distinguish well between (b) and (c).

In order to develop a discourse-aware evaluation metric, we first generate discourse trees for the reference and the system-translated sentences using a discourse parser, and then we measure the similarity between the two discourse trees. We describe these two steps below.

3.1 Generating Discourse Trees

In Rhetorical Structure Theory, discourse analysis involves two subtasks: (*i*) *discourse segmentation*, or breaking the text into a sequence of EDUs, and (*ii*) *discourse parsing*, or the task of linking the units (EDUs and larger discourse units) into labeled discourse trees. Recently, Joty et al. (2012) proposed discriminative models for both discourse segmentation and discourse parsing at the sentence level. The segmenter uses a maximum entropy model that achieves state-of-the-art accuracy on this task, having an F_1-score of 90.5%, while human agreement is 98.3%.

The discourse parser uses a dynamic Conditional Random Field (Sutton et al., 2007) as a parsing model in order to infer the probability of all possible discourse tree constituents. The inferred (posterior) probabilities are then used in a probabilistic CKY-like bottom-up parsing algorithm to find the most likely DT. Using the standard set of 18 coarse-grained relations defined in (Carlson and Marcu, 2001), the parser achieved an F_1-score of 79.8%, which is very close to the human agreement of 83%. These high scores allowed us to develop successful discourse similarity metrics.[2]

3.2 Measuring Similarity

A number of metrics have been proposed to measure the similarity between two labeled trees, e.g., Tree Edit Distance (Tai, 1979) and Tree Kernels (Collins and Duffy, 2001; Moschitti and Basili, 2006). Tree kernels (TKs) provide an effective way to integrate arbitrary tree structures in kernel-based machine learning algorithms like SVMs.

In the present work, we use the convolution TK defined in (Collins and Duffy, 2001), which efficiently calculates the number of common subtrees in two trees. Note that this kernel was originally designed for syntactic parsing, where the subtrees are subject to the constraint that their nodes are taken with either all or none of the children. This constraint of the TK imposes some limitations on the type of substructures that can be compared.

[2]The discourse parser is freely available from http://alt.qcri.org/tools/

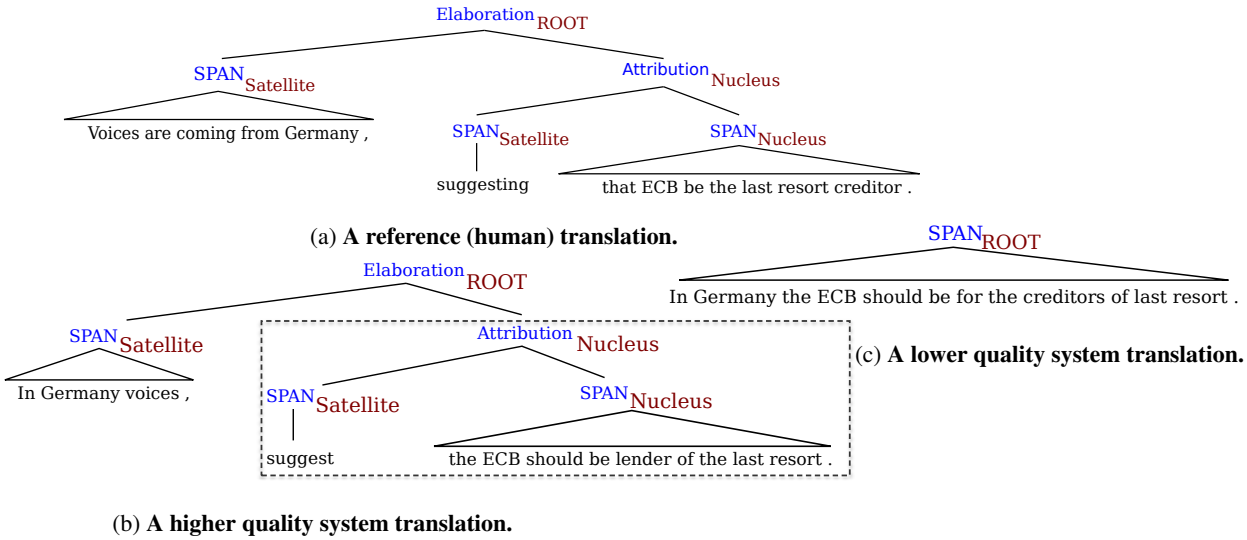


Figure 1: Example of three different discourse trees for the translations of a source sentence. (a) The reference, (b) A higher quality translation, (c) A lower quality translation.

One way to cope with the limitations of the TK is to change the representation of the trees to a form that is suitable to capture the relevant information for our task. We experiment with TKs applied to two different representations of the discourse tree: non-lexicalized (DR), and lexicalized (DR-LEX). In Figure 2 we show the two representations for the subtree that spans the text: *"suggest the ECB should be the lender of last resort"*, which is highlighted in Figure 1b.

As shown in Figure 2a, DR does not include any lexical item, and therefore measures the similarity between two translations in terms of their discourse structures only. On the contrary, DR-LEX includes the lexical items to account for lexical matching; moreover, it separates the structure (the skeleton) of the tree from its labels, i.e. the nuclearity and the relations, in order to allow the tree kernel to give partial credit to subtrees that differ in labels but match in their skeletons. More specifically, it uses the tags SPAN and EDU to build the skeleton of the tree, and considers the nuclearity and/or the relation labels as properties, added as children, of these tags.

For example, a SPAN has two properties (its nuclearity and its relation), and an EDU has one property (its nuclearity). The words of an EDU are placed under the predefined children NGRAM. In order to allow the tree kernel to find subtree matches at the word level, we include an additional layer of *dummy* leaves as was done in (Moschitti et al., 2007); not shown in Figure 2, for simplicity.

4 Experimental Setup

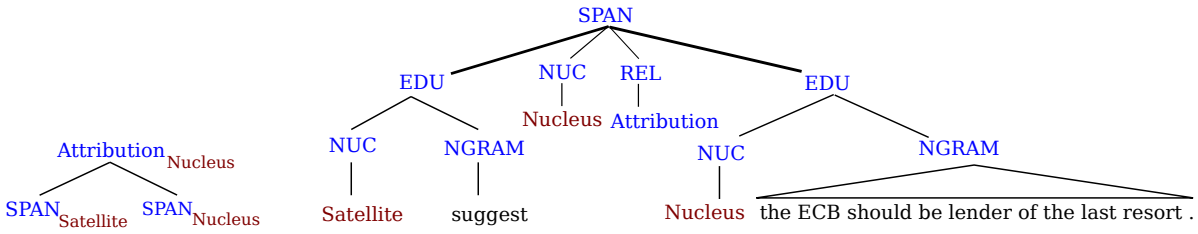
In our experiments, we used the data available for the WMT12 and the WMT11 metrics shared tasks for translations into English.[3] This included the output from the systems that participated in the WMT12 and the WMT11 MT evaluation campaigns, both consisting of 3,003 sentences, for four different language pairs: Czech-English (CS-EN), French-English (FR-EN), German-English (DE-EN), and Spanish-English (ES-EN); as well as a dataset with the English references.

We measured the correlation of the metrics with the human judgments provided by the organizers. The judgments represent rankings of the output of five systems chosen at random, for a particular sentence, also chosen at random. Note that each judgment effectively constitutes 10 pairwise system rankings. The overall coverage, i.e. the number of unique sentences that were evaluated, was only a fraction of the total; the total number of judgments, along with other information of the datasets are shown in Table 1.

4.1 MT Evaluation Metrics

In this study, we evaluate to what extent existing evaluation metrics can benefit from additional discourse information. To do so, we contrast different MT evaluation metrics with and without discourse information. The evaluation metrics we used are described below.

[3] http://www.statmt.org/wmt{11,12}/results.html



(a) DT for DR (b) DT for DR-LEX

Figure 2: Two different DT representations for the highlighted subtree shown in Figure 1b.

	WMT12				WMT11			
	systs	ranks	sents	judges	systs	ranks	sents	judges
CS-EN	6	1,294	951	45	8	498	171	20
DE-EN	16	1,427	975	47	20	924	303	31
ES-EN	12	1,141	923	45	15	570	207	18
FR-EN	15	1,395	949	44	18	708	249	32

Table 1: Number of systems (systs), judgments (ranks), unique sentences (sents), and different judges (judges) for the different language pairs, for the human evaluation of the WMT12 and WMT11 shared tasks.

Metrics from WMT12. We used the publicly available scores for all metrics that participated in the WMT12 metrics task (Callison-Burch et al., 2012): SPEDE07PP, AMBER, METEOR, TERRORCAT, SIMPBLEU, XENERRCATS, WORDBLOCKEC, BLOCKERRCATS, and POSF.

Metrics from ASIYA. We used the freely available version of the ASIYA toolkit[4] in order to extend the set of evaluation measures contrasted in this study beyond those from the WMT12 metrics task. ASIYA (Giménez and Màrquez, 2010a) is a suite for MT evaluation that provides a large set of metrics that use different levels of linguistic information. For reproducibility, below we explain the individual metrics with the exact names required by the toolkit to calculate them.

First, we used ASIYA's ULC (Giménez and Màrquez, 2010b), which was the best performing metric at the system and the segment levels at the WMT08 and WMT09 metrics tasks. This is a uniform linear combination of 12 individual metrics. From the original ULC, we only replaced TER and Meteor individual metrics by newer versions taking into account synonymy lookup and paraphrasing: TERp-A and METEOR-pa in ASIYA's terminology. We will call this combined metric Asiya-0809 in our experiments.

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

To complement the set of individual metrics that participated at the WMT12 metrics task, we also computed the scores of other commonly-used evaluation metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002), TER (Snover et al., 2006), ROUGE-W (Lin, 2004), and three METEOR variants (Denkowski and Lavie, 2011): METEOR-ex (exact match), METEOR-st (+stemming) and METEOR-sy (+synonyms). The uniform linear combination of the previous 7 individual metrics plus the 12 from Asiya-0809 is reported as Asiya-ALL in the experimental section.

The individual metrics combined in Asiya-ALL can be naturally categorized according to the type of linguistic information they use to compute the quality scores. We grouped them in the following four families and calculated the uniform linear combination of the metrics in each group:[5]

1. Asiya-LEX. Combination of five metrics based on lexical similarity: BLEU, NIST, METEOR-ex, ROUGE-W, and TERp-A.
2. Asiya-SYN. Combination of four metrics ba-sed on syntactic information from constituency and dependency parse trees: 'CP-STM-4', 'DP-HWCM_c-4', 'DP-HWCM_r-4', and 'DP-Or(*)'.
3. Asiya-SRL. Combination of three metric variants based on predicate argument structures (semantic role labeling): 'SR-Mr(*)', 'SR-Or(*)', and 'SR-Or'.
4. Asiya-SEM. Combination of two metrics variants based on semantic parsing:[6] 'DR-Or(*)' and 'DR-Orp(*)'.

[5]A detailed description of every individual metric can be found at (Giménez and Màrquez, 2010b). For a more up-to-date description, see the User Manual from ASIYA's website.

[6]In ASIYA the metrics from this family are referred to as "Discourse Representation" metrics. However, the structures they consider are actually very different from the discourse structures exploited in this paper. See the discussion in Section 2. For clarity, we will refer to them as *semantic parsing* metrics.

All uniform linear combinations are calculated outside ASIYA. In order to make the scores of the different metrics comparable, we performed a $\min$–$\max$ normalization, for each metric, and for each language pair combination.

4.2 Human Judgements and Learning

The human-annotated data from the WMT campaigns encompasses series of rankings on the output of different MT systems for every source sentence. Annotators rank the output of five systems according to perceived translation quality. The organizers relied on a random selection of systems, and a large number of comparisons between pairs of them, to make comparisons across systems feasible (Callison-Burch et al., 2012). As a result, for each source sentence, only relative rankings were available. As in the WMT12 experimental setup, we use these rankings to calculate correlation with human judgments at the sentence-level, i.e. Kendall's Tau; see (Callison-Burch et al., 2012) for details.

For the experiments reported in Section 5.4, we used pairwise rankings to discriminatively learn the weights of the linear combinations of individual metrics. In order to use the WMT12 data for training a learning-to-rank model, we transformed the five-way relative rankings into ten pairwise comparisons. For instance, if a judge ranked the output of systems A, B, C, D, E as $A > B > C > D > E$, this would entail that $A > B$, $A > C$, $A > D$ and $A > E$, etc.

To determine the relative weights for the tuned combinations, we followed a similar approach to the one used by PRO to tune the relative weights of the components of a log-linear SMT model (Hopkins and May, 2011), also using Maximum Entropy as the base learning algorithm. Unlike PRO, (*i*) we use *human judgments*, not automatic scores, and (*ii*) we train on *all pairs*, not on a subsample.

5 Experimental Results

In this section, we explore how discourse information can be used to improve machine translation evaluation metrics. Below we present the evaluation results at the system- and segment-level, using our two basic metrics on discourse trees (Section 3.1), which are referred to as DR and DR-LEX.

5.1 Evaluation

In our experiments, we only consider translation into English, and use the data described in Table 1. For evaluation, we follow the setup of the metrics task of WMT12 (Callison-Burch et al., 2012): at the system-level, we use the official script from WMT12 to calculate the Spearman's correlation, where higher *absolute* values indicate better metrics performance; at the segment-level, we use Kendall's Tau for measuring correlation, where negative values are worse than positive ones.[7]

In our experiments, we combine DR and DR-LEX to other metrics in two different ways: using uniform linear interpolation (at system- and segment-level), and using a tuned linear interpolation for the segment-level. We only present the average results over all four language pairs. For simplicity, in our tables we show results divided into evaluation groups:

1. Group I: contains our evaluation metrics, DR and DR-LEX.

2. Group II: includes the metrics that participated in the WMT12 metrics task, excluding metrics which did not have results for all language pairs.

3. Group III: contains other important evaluation metrics, which were not considered in the WMT12 metrics task: NIST and ROUGE for both system- and segment-level, and BLEU and TER at segment-level.

4. Group IV: includes the metric combinations calculated with ASIYA and described in Section 4.

For each metric in groups II, III and IV, we present the results for the original metric as well for the linear interpolation of that metric with DR and with DR-LEX. The combinations with DR and DR-LEX that improve over the original metrics are shown in **bold**, and those that degrade are in *italic*. Furthermore, we also present overall results for: (*i*) the average score over all metrics, excluding DR and DR-LEX, and (*ii*) the differences in the correlations for the DR/DR-LEX-combined and the original metrics.

[7] We have fixed a bug in the scoring tool from WMT12, which was making all scores positive. This made TERRORCAT's score negative, as we present it in Table 3.

	Metrics		+DR	+DR-LEX
I	DR	.807	–	–
	DR-LEX	.876	–	–
II	SEMPOS	.902	*.853*	**.903**
	AMBER	.857	*.829*	**.869**
	METEOR	.834	**.861**	**.888**
	TERRORCAT	.831	**.854**	**.889**
	SIMPBLEU	.823	**.826**	**.859**
	TER	.812	**.836**	**.848**
	BLEU	.810	**.830**	**.846**
	POSF	.754	**.841**	**.857**
	BLOCKERRCATS	.751	**.859**	**.855**
	WORDBLOCKEC	.738	**.822**	**.843**
	XENERRCATS	.735	**.819**	**.843**
III	NIST	.817	**.842**	**.875**
	ROUGE	.884	**.899**	*.869*
IV	Asiya-LEX	.879	**.881**	**.882**
	Asiya-SYN	.891	**.913**	*.883*
	Asiya-SRL	.917	*.911*	*.909*
	Asiya-SEM	.891	*.889*	*.886*
	Asiya-0809	.905	**.914**	.905
	Asiya-ALL	.899	**.907**	*.896*
	average	.839	**.862**	**.874**
	diff.		**+.024**	**+.035**

Table 2: Results on WMT12 at the system-level. Spearman's correlation with human judgments.

	Metrics		+DR	+DR-LEX
I	DR	-.433	–	–
	DR-LEX	.133	–	–
II	SPEDE07PP	.254	*.190*	*.223*
	METEOR	.247	*.178*	*.217*
	AMBER	.229	*.180*	*.216*
	SIMPBLEU	.172	*.141*	**.191**
	XENERRCATS	.165	*.132*	**.185**
	POSF	.154	*.125*	**.201**
	WORDBLOCKEC	.153	*.122*	**.181**
	BLOCKERRCATS	.074	*.068*	**.151**
	TERRORCAT	-.186	**-.111**	**-.104**
III	NIST	.214	*.172*	*.206*
	ROUGE	.185	*.144*	**.201**
	TER	.217	*.179*	**.229**
	BLEU	.185	*.154*	**.190**
IV	Asiya-LEX	.254	*.237*	*.253*
	Asiya-SYN	.177	*.169*	**.191**
	Asiya-SRL	-.023	**.015**	**.161**
	Asiya-SEM	.134	**.152**	**.197**
	Asiya-0809	.254	*.250*	**.258**
	Asiya-ALL	.268	*.265*	**.270**
	average	.165	*.145*	**.190**
	diff.		*-.019*	**+.026**

Table 3: Results on WMT12 at the segment-level. Kendall's Tau with human judgments.

5.2 System-level Results

Table 2 shows the system-level experimental results for WMT12. We can see that DR is already competitive by itself: on average, it has a correlation of .807, very close to BLEU and TER scores (.810 and .812, respectively). Moreover, DR yields improvements when combined with 15 of the 19 metrics; worsening only four of the metrics. Overall, we observe an average improvement of +.024, in the correlation with the human judgments. This suggests that DR contains information that is complementary to that used by the other metrics. Note that this is true both for the individual metrics from groups II and III, as well as for the metric combinations in group IV. Combinations in the last group involve several metrics that already use linguistic information at different levels and are hard to improve over; yet, adding DR does improve, which shows that it has some complementary information to offer.

As expected, DR-LEX performs better than DR since it is lexicalized (at the unigram level), and also gives partial credit to correct structures. Individually, DR-LEX outperforms most of the metrics from group II, and ranks as the second best metric in that group. Furthermore, when combined with individual metrics in group II, DR-LEX is able to improve consistently over each one of them.

Note that, even though DR-LEX has better individual performance than DR, it does not yield improvements when combined with most of the metrics in group IV.[8] However, over all metrics and all language pairs, DR-LEX is able to obtain an average improvement in correlation of +.035, which is remarkably higher than that of DR. Thus, we can conclude that at the system-level, adding discourse information to a metric, even using the simplest of the combination schemes, is a good idea for most of the metrics, and can help to significantly improve the correlation with human judgments.

5.3 Segment-level Results: Non-tuned

Table 3 shows the results for WMT12 at the segment-level. We can see that DR performs badly, with a high negative Kendall's Tau of -.433. This should not be surprising: (a) the discourse tree structure alone does not contain enough information for a good evaluation at the segment-level, and (b) this metric is more sensitive to the quality of the DT, which can be wrong or void.

[8] In this work, we have not investigated the reasons behind this phenomenon. We speculate that this might be caused by the fact that the lexical information in DR-LEX is incorporated only in the form of unigram matching at the sentence-level, while the metrics in group IV are already complex combined metrics, which take into account stronger lexical models. Note, however, that the variations are very small and might not be significant.

	Metrics	Orig.	Tuned	+DR	+DR-LEX
I	DR	-.433	–	–	–
	DR-LEX	.133	–	–	–
II	SPEDE07PP	.254	–	*.253*	.254
	METEOR	.247	–	**.250**	**.251**
	AMBER	.229	–	**.230**	**.232**
	SIMPBLEU	.172	–	**.181**	**.199**
	TERRORCAT	-.186	–	**.181**	**.196**
	XENERRCATS	.165	–	**.175**	**.194**
	POSF	.154	–	**.160**	**.201**
	WORDBLOCKEC	.153	–	**.161**	**.189**
	BLOCKERRCATS	.074	–	**.087**	**.150**
III	NIST	.214	–	**.222**	**.224**
	ROUGE	.185	–	**.196**	**.218**
	TER	.217	–	**.229**	**.246**
	BLEU	.185	–	**.189**	**.194**
IV	Asiya-LEX	.254	.266	**.269**	**.270**
	Asiya-SYN	.177	.229	.228	**.232**
	Asiya-SRL	-.023	-.004	**.039**	**.181**
	Asiya-SEM	.134	.146	**.179**	**.202**
	Asiya-0809	.254	.295	.295	.295
	Asiya-ALL	.268	.296	.295	.295
	average	.165		.201	.222
	diff.			**+.036**	**+.057**

Table 4: Results on WMT12 at the segment-level: tuning with cross-validation on WMT12. Kendall's Tau with human judgments.

Additionally, DR is more likely to produce a high number of ties, which is harshly penalized by WMT12's definition of Kendall's Tau. Conversely, ties and incomplete discourse analysis were not a problem at the system-level, where evidence from all 3,003 test sentences is aggregated, and allows to rank systems more precisely. Due to the low score of DR as an individual metric, it fails to yield improvements when uniformly combined with other metrics.

Again, DR-LEX is better than DR; with a positive Tau of +.133, yet as an individual metric, it ranks poorly compared to other metrics in group II. However, when linearly combined with other metrics, DR-LEX outperforms 14 of the 19 metrics in Table 3. Across all metrics, DR-LEX yields an average Tau improvement of +.026, i.e. from .165 to .190. This is a large improvement, taking into account that the combinations are just uniform linear combinations. In subsection 5.4, we present the results of tuning the linear combination in a discriminative way.

5.4 Segment-level Results: Tuned

We experimented with tuning the weights of the individual metrics in the metric combinations, using the learning method described in Section 4.2. First, we did this using cross-validation to tune and test on WMT12. Later we tuned on WMT12 and evaluated on WMT11. For cross-validation in WMT12, we used ten folds of approximately equal sizes, each containing about 300 sentences: we constructed the folds by putting together entire documents, thus not allowing sentences from a document to be split over two different folds. During each cross-validation run, we trained our pairwise ranker using the human judgments corresponding to nine of the ten folds. We aggregated the data for different language pairs, and produced a single set of tuning weights for all language pairs.[9] We then used the remaining fold for evaluation

The results are shown in Table 4. As in previous sections we present the average results over all four language pairs. We can see that the tuned combinations with DR-LEX improve over most of the individual metrics in groups II and III. Interestingly, the tuned combinations that include the much weaker metric DR now improve over 12 out of 13 of the individual metrics in groups II and III, and only slightly degrades the score of the 13th one (SPEDE07PP).

Note that the ASIYA metrics are combinations of several metrics, and these combinations (which exclude DR and DR-LEX) can be also tuned; this yields sizable improvements over the untuned versions as column three in the table shows. Compared to this baseline, DR improves for three of the six ASIYA metrics, while DR-LEX improves for four of them. Note that improving over the last two ASIYA metrics is very hard: they have very high scores of .296 and .295; for comparison, the best segment-level system at WMT12 (SPEDE07PP) achieved a Tau of .254.

On average, DR improves Tau from .165 to .201, which is +.036, while DR-LEX improves to .222, or +.057. These much larger improvements highlight the importance of tuning the linear combination when working at the segment-level.

5.4.1 Testing on WMT11

In order to rule out the possibility that the improvement of the tuned metrics on WMT12 comes from over-fitting, and to verify that the tuned metrics do generalize when applied to other sentences, we also tested on a new test set: WMT11.

[9] Tuning separately for each language pair yielded slightly lower results.

Therefore, we tuned the weights on *all* WMT12 pairwise judgments (no cross-validation), and we evaluated on WMT11. Since the metrics that participated in WMT11 and WMT12 are different (and even when they have the same name, there is no guarantee that they have not changed from 2011 to 2012), we only report results for the versions of NIST, ROUGE, TER, and BLEU available in ASIYA, as well as for the ASIYA metrics, thus ensuring that the metrics in the experiments are consistent for 2011 and 2012.

The results are shown in Table 5. Once again, tuning yields sizable improvements over the simple combination for the ASIYA metrics (third column in Table 5). Adding DR and DR-LEX to the combinations manages to improve over five and four of the six tuned ASIYA metrics, respectively. However, some of the differences are very small. On the contrary, DR and DR-LEX significantly improve over NIST, ROUGE, TER, and BLEU. Overall, DR improves the average Tau from .207 to .244, which is +.037, while DR-LEX improves to .267 or +.061. These improvements are very close to those for the WMT12 cross-validation. This shows that the weights learned on WMT12 generalize well, as they are also good for WMT11.

What is also interesting to note is that when tuning is used, DR helps achieve sizeable improvements, even if not as strong as for DR-LEX. This is remarkable given that DR has a strong negative Tau as an individual metric at the sentence-level. This suggests that both DR and DR-LEX contain information that is complementary to that of the individual metrics that we experimented with.

Overall, from the experimental results in this section, we can conclude that discourse structure is an important information source to be taken into account in the automatic evaluation of machine translation output.

			Tuned		
	Metrics	**Orig.**		+DR	+DR-LEX
I	DR	-.447	–	–	–
	DR-LEX	.146	–	–	–
III	NIST	.219	–	**.226**	**.232**
	ROUGE	.205	–	**.218**	**.242**
	TER	.262	–	**.274**	**.296**
	BLEU	.186	–	**.192**	**.207**
IV	Asiya-LEX	.282	.301	**.302**	**.303**
	Asiya-SYN	.216	.259	**.260**	**.260**
	Asiya-SRL	-.004	.017	**.051**	**.200**
	Asiya-SEM	.189	.194	**.220**	**.239**
	Asiya-0809	.300	.348	**.349**	.348
	Asiya-ALL	.313	.347	.347	.347
	average	.207		.244	.267
	diff.			**+.037**	**+.061**

Table 5: Results on WMT11 at the segment-level: tuning on the entire WMT12. Kendall's Tau with human judgments.

6 Conclusions and Future Work

In this paper we have shown that discourse structure can be used to improve automatic MT evaluation. First, we defined two simple discourse-aware similarity metrics (lexicalized and un-lexicalized), which use the all-subtree kernel to compute similarity between discourse parse trees in accordance with the Rhetorical Structure Theory. Then, after extensive experimentation on WMT12 and WMT11 data, we showed that a variety of existing evaluation metrics can benefit from our discourse-based metrics, both at the segment- and the system-level, especially when the discourse information is incorporated in an informed way (i.e. using supervised tuning). Our results show that discourse-based metrics can improve the state-of-the-art MT metrics, by increasing correlation with human judgments, even when only sentence-level discourse information is used.

Addressing discourse-level phenomena in MT is a relatively new research direction. Yet, many of the ongoing efforts have been moderately successful according to traditional evaluation metrics. There is a consensus in the MT community that more discourse-aware metrics need to be proposed for this area to move forward. We believe this work is a valuable contribution towards this longer-term goal.

The tuned combined metrics tested in this paper are just an initial proposal, i.e. a simple adjustment of the relative weights for the individual metrics in a linear combination. In the future, we plan to work on integrated representations of syntactic, semantic and discourse-based structures, which would allow us to train evaluation metrics based on more fine-grained features. Additionally, we propose to use the discourse information for MT in two different ways. First, at the sentence-level, we can use discourse information to re-rank alternative MT hypotheses; this could be applied either for MT parameter tuning, or as a post-processing step for the MT output. Second, we propose to move in the direction of using discourse information beyond the sentence-level.

References

Nicholas Asher and Alex Lascarides, 2003. *Logics of Conversation*. Cambridge University Press.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July. ACL.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June. ACL.

Lynn Carlson and Daniel Marcu. 2001. Discourse Tagging Reference Manual. Technical Report ISI-TR-545, University of Southern California Information Sciences Institute.

Bruno Cartoni, Sandrine Zufferey, Thomas Meyer, and Andrei Popescu-Belis. 2011. How comparable are parallel corpora? measuring the distribution of general vocabulary and connectives. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 78–86, Portland, Oregon, June. ACL.

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

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 263–270, Ann Arbor, Michigan.

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

Elisabet Comelles, Jesús Giménez, Lluís Màrquez, Irene Castellón, and Victoria Arranz. 2010. Document-level automatic mt evaluation based on discourse representations. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 333–338, Uppsala, Sweden, July. ACL.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland, July. ACL.

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. Morgan Kaufmann Publishers Inc.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology*, HLT-NAACL, pages 273–280.

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*, pages 256–264, Prague, Czech Republic, June. ACL.

Jesús Giménez and Lluís Màrquez. 2009. On the robustness of syntactic and semantic features for automatic MT evaluation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 250–258, Athens, Greece, March. ACL.

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(3–4):77–86.

Michael Halliday and Ruqaiya Hasan, 1976. *Cohesion in English*. Longman, London.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 283–289.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1179–1190, Jeju Island, Korea. ACL.

Christian Hardmeier. 2012. Discourse in statistical machine translation. a survey and a case study. *Discours. Revue de linguistique, psycholinguistique et informatique*, 11(8726).

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2012. A Novel Discriminative Framework for Sentence-Level Discourse Analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 904–915, Jeju Island, Korea. ACL.

Shafiq Joty, Giuseppe Carenini, Raymond T. Ng, and Yashar Mehdad. 2013. Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, pages 486–496, Sofia, Bulgaria. ACL.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Model theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory.* Number 42 in Studies in Linguistics and Philosophy. Kluwer Academic Publishers.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of Workshop on Text Summarization Branches Out*, pages 74–81, Barcelona.

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, Ann Arbor, Michigan, June. ACL.

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, Montréal, Canada, June. ACL.

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

Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. 2012. Machine translation of labeled discourse connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*.

Thomas Meyer. 2011. Disambiguating temporal-contrastive connectives for machine translation. In *Proceedings of the ACL 2011 Student Session*, pages 46–51, Portland, OR, USA, June. ACL.

Alessandro Moschitti and Roberto Basili. 2006. A Tree Kernel approach to Question and Answer Classification in Question Answering Systems. In *Proceedings of the 5th international conference on Language Resources and Evaluation*, Genoa, Italy.

Alessandro Moschitti, Silvia Quarteroni, Roberto Basili, and Suresh Manandhar. 2007. Exploiting Syntactic and Shallow Semantic Kernels for Question/Answer Classification. In *Proceedings of the ACL-2007*, pages 776–783, Prague, Czech Republic.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics (ACL'02)*, Philadelphia, PA, USA.

Maja Popovic 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*, pages 48–55, Prague, Czech Republic, June. ACL.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal smt. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 271–279, Ann Arbor, Michigan.

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 7th Biennial Conference of the Association for Machine Translation in the Americas*, AMTA '06, Cambridge, MA, USA.

Charles Sutton, Andrew McCallum, and Khashayar Rohanimanesh. 2007. Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data. *Journal of Machine Learning Research (JMLR)*, 8:693–723.

Kuo-Chung Tai. 1979. The tree-to-tree correction problem. *Journal of the ACM*, 26(3):422–433, July.

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

Bonnie Webber, Andrei Popescu-Belis, Katja Markert, and Jörg Tiedemann, editors. 2013. *Proceedings of the Workshop on Discourse in Machine Translation*. ACL, Sofia, Bulgaria, August.

Bonnie Webber. 2004. D-LTAG: Extending Lexicalized TAG to Discourse. *Cognitive Science*, 28(5):751–779.

Billy T. M. Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL, pages 1060–1068, Jeju Island, Korea, July. ACL.

Dekai Wu and Pascale Fung. 2009. Semantic roles for smt: A hybrid two-pass model. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 13–16, Boulder, Colorado, June.