# Discourse Structure in Machine Translation Evaluation

Shafiq Joty[*]
School of Computer Science and
Engineering
Nanyang Technological University

Francisco Guzmán[**]
Applied Machine Learning Group
Facebook

Lluís Màrquez[†]
ALT, QCRI, HBKU, Qatar Foundation

Preslav Nakov[†]
ALT, QCRI, HBKU, Qatar Foundation

*In this article, we explore the potential of using sentence-level discourse structure for machine translation evaluation. We first design discourse-aware similarity measures, which use all-subtree kernels to compare discourse parse trees in accordance with the Rhetorical Structure Theory (RST). Then, we show that a simple linear combination with these measures can help improve various existing machine translation evaluation metrics regarding correlation with human judgments both at the segment level and at the system level. This suggests that discourse information is complementary to the information used by many of the existing evaluation metrics, and thus it could be taken into account when developing richer evaluation metrics, such as the WMT-14 winning combined metric* DISCOTK$_{party}$. *We also provide a detailed analysis of the relevance of various discourse elements and relations from the RST parse trees for machine translation evaluation. In particular, we show that (i) all aspects of the RST tree are relevant, (ii) nuclearity is more useful than relation type, and (iii) the similarity of the translation RST tree to the reference RST tree is positively correlated with translation quality.*

[*] School of Computer Science and Engineering, Nanyang Technological University. E-mail: `srjoty@ntu.edu.sg`.

[**] Applied Machine Learning Group, Facebook. E-mail: `fguzman@fb.com`.

[†] HBKU Research Complex B1, P.O. Box 5825, Doha, Qatar. E-mail: {`lmarquez,pnakov`}`@qf.org.qa`.

Submission received: 26 August 2016; revised version received: 12 December 2016; accepted for publication: 21 April 2017.

doi:10.1162/COLI_a_00298

## 1. Introduction

From its foundations, Statistical Machine Translation (SMT) as a field 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* (Brown et al. 1993; Koehn, Och, and Marcu 2003).

Although modern SMT systems switched to a discriminative log-linear framework (Och 2003; Watanabe et al. 2007; Chiang, Marton, and Resnik 2008; Hopkins and May 2011), 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.

One of the fruitful research directions for improving SMT has been the usage of more structured linguistic information. For instance, in SMT we find systems based on syntax (Galley et al. 2004; Quirk, Menezes, and Cherry 2005), hierarchical structures (Chiang 2005), and semantic roles (Wu and Fung 2009; Lo, Tumuluru, and Wu 2012; Bazrafshan and Gildea 2014). However, it was not until recently that syntax-based SMT systems started to outperform their phrase-based counterparts, especially for language pairs that need long-distance reordering such as Chinese–English and German–English (Nadejde, Williams, and Koehn 2013).

Another less-explored way consists of going beyond the sentence-level; for example, translating at the document level or taking into account broader contextual information. The idea is to obtain adequate translations respecting cross-sentence relations and enforcing cohesion and consistency at the document level (Hardmeier, Nivre, and Tiedemann 2012; Ben et al. 2013; Louis and Webber 2014; Tu, Zhou, and Zong 2014; Xiong, Zhang, and Wang 2015). Research in this direction has also been the focus of the two editions of the DiscoMT workshop, in 2013 and 2015 (Webber et al. 2013, 2015; Hardmeier et al. 2015).

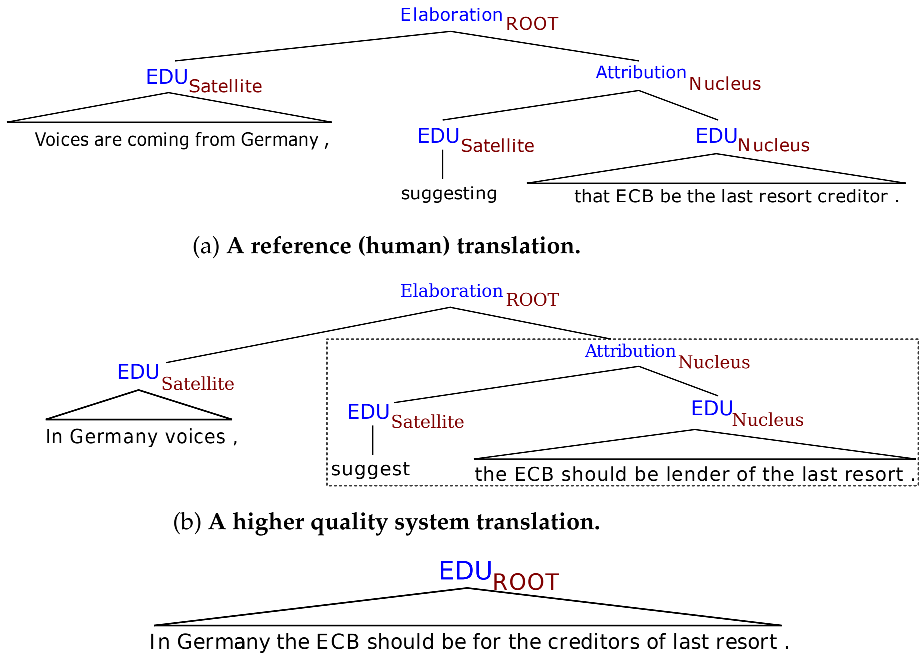
*Automatic MT evaluation* is an integral part of the process of developing and tuning an SMT system. Reference-based evaluation measures compare the output of a system to one or more human translations (called **references**) and produce a similarity score indicating the quality of the translation. The first metrics approached similarity as a shallow word *n*-gram matching between the translation and one or more references, with a limited use of linguistic information. BLEU (Papineni et al. 2002) is the best-known metric in this family, and has been used for years as the evaluation standard in the MT community. BLEU can be efficiently calculated and has shown good correlation with human assessments when evaluating systems on large quantities of text. However, it is also known that BLEU and similar metrics are unreliable for high-quality translation output (Doddington 2002; Lavie and Agarwal 2007), and they cannot tell apart raw machine translation output from a fully fluent professionally post-edited version thereof (Denkowski and Lavie 2012). Moreover, lexical-matching similarity has been shown to be both insufficient and not strictly necessary for two sentences to convey the same meaning (Coughlin 2003; Culy and Riehemann 2003; Callison-Burch, Osborne, and Koehn 2006).

Several alternatives emerged to overcome these limitations, most notably TER (Snover et al. 2006) and METEOR (Lavie and Denkowski 2009). Researchers have explored, with good results, the addition of other levels of linguistic information, including synonymy and paraphrasing (Lavie and Denkowski 2009), syntax (Liu and Gildea 2005; Giménez and Màrquez 2007; Popovic and Ney 2007), semantic roles (Giménez and Màrquez 2007; Lo, Tumuluru, and Wu 2012), and, most recently, discourse (Giménez et al. 2010; Wong and Kit 2012; Guzmán et al. 2014a, 2014b; Joty et al. 2014).

Beyond all previous considerations, MT systems are usually evaluated by computing translation quality on individual sentences and performing some simple aggregation to produce the *system-level* evaluation scores. To the best of our knowledge, semantic relations between clauses in a sentence and between sentences in a text have not been seriously explored. However, clauses and sentences rarely stand on their own in a well-written text; rather, the logical relationship between them carries significant information that allows the text to express a meaning as a whole. Each clause follows smoothly from the ones before it and leads into the ones that come afterward. This logical relationship between clauses forms a **coherence structure** (Hobbs 1979). In discourse analysis, we seek 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). Rhetorical Structure Theory (RST; Mann and Thompson 1988) is perhaps the most influential of these in computational linguistics, where it is used either to parse the text in language understanding or to plan a coherent text in language generation (Taboada and Mann 2006). RST describes coherence using *discourse relations* between parts of a text and postulates a hierarchical tree structure called *discourse tree*. For example, Figure 1 in the next section shows discourse trees for three different translations of a source sentence.

Modeling discourse brings together the usage of higher-level linguistic information and the exploration of relations between clauses and sentences in a text, which makes it a very attractive goal for MT and its evaluation. We believe that the semantic and pragmatic information captured in the form of discourse trees (i) can yield better



(a) **A reference (human) translation.**

(b) **A higher quality system translation.**

(c) **A lower quality system translation.**

**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.

MT evaluation metrics, and (ii) can help develop discourse-aware SMT systems that produce more coherent translations.

In this work, we focus on the first of the two previous research hypotheses. Specifically, we show that sentence-level discourse information can be used to produce reference-based evaluation measures that perform well on their own, but more importantly, can be used to improve over many existing MT evaluation metrics regarding correlation with human assessments. We conduct our research in three steps. First, we design a simple discourse-aware similarity measure, DR-LEX, based on RST trees, generated with a publicly available discourse parser (Joty, Carenini, and Ng 2015), and the well-known **all subtree** kernel (Collins and Duffy 2001). The subtree kernel computes a similarity value by comparing the discourse tree representation of a system translation with that of a reference translation. We show that a simple uniform linear combination with this metric helps to improve a large number of MT evaluation metrics at the segment-level and at the system-level in the context of the WMT11 and the WMT12 metrics shared task benchmarks (Callison-Burch et al. 2011, 2012). Second, we show that tuning (i.e., learning) the weights in the linear combination of metrics using human-assessed examples is a robust way to improve the effectiveness of the DR-LEX metric significantly. Following the idea of an interpolated combination, we put together several variants of our discourse metric (using different tree-based representations) with many strong pre-existing metrics provided by the ASIYA toolkit for MT evaluation (Gonzàlez, Giménez, and Màrquez 2012). The result is DISCOTK$_{party}$, which scored best at the WMT14 Metrics task (Bojar et al. 2014), both at the system level and at the segment level. Third, we conduct an ablation study that helps us understand which elements of the discourse parse tree have the highest impact on the quality of the evaluation measure. Interestingly enough, the **nuclearity** feature (i.e., the distinction between main and subordinate units) of the RST tree turns out to be more important than the discourse relation types (e.g., *Elaboration*, *Contrast*).

Note that, although extensive, this study is restricted to sentence-level evaluation, which arguably can limit the benefits of using global discourse properties (i.e., document-level discourse structure). Fortunately, many sentences are long and complex enough to present rich discourse structures connecting their basic clauses. Thus, although limited, this setting can demonstrate the potential of discourse-level information for MT evaluation. Furthermore, sentence-level scoring is compatible with most translation systems, which work on a sentence-by-sentence basis. It could also be beneficial to modern MT tuning mechanisms such as PRO (Hopkins and May 2011) and MIRA (Watanabe et al. 2007; Chiang, Marton, and Resnik 2008), which also work at the sentence level. Finally, it could also be used for re-ranking $n$-best lists of translation hypotheses.

The rest of the paper is organized as follows. Section 2 introduces our proposal for a family of discourse-based similarity metrics. Sections 3 and 4 describe the experimental setting and the evaluation of the discourse-based metrics, alone and in combination with other pre-existing measures. Section 5 empirically analyzes the main discourse-based metric and performs an ablation study to better understand its contributions. Finally, Sections 6 and 7 discuss related work and present the conclusions together with some directions for future research.

## 2. Discourse-Based Similarity Measures

Different formal theories of discourse have been proposed in the literature, reflecting different viewpoints about what is the best way to describe the coherence structure of a text. For example, Asher and Lascarides (2003) proposed the Segmented Discourse Representation Theory, which is driven by sentence semantics. Webber (2004) and Danlos (2009) extended the sentence grammar to formalize discourse structure. Mann and Thompson (1988) proposed RST, which was inspired by empirical analysis of authentic texts. Although RST was initially intended to be used for text generation, it later became popular as a framework for parsing the structure of a text. This work relies on RST-based coherence structure.

RST posits a tree representation of a text, known as a discourse tree. As shown in Figure 1(a), the leaves of a discourse tree (three in this example) correspond to contiguous atomic clause-like text spans, called **elementary discourse units (EDUs)**, which serve as building blocks for constructing the tree. In the tree, adjacent EDUs are connected by certain **coherence relations** (e.g., *Elaboration*, *Attribution*), thus forming larger discourse units, which in turn are also subject to this process of relation-based linking. Discourse units that are linked by a relation are further distinguished based on their relative importance in the text: **nuclei** are the core arguments of the relation, and **satellites** are supportive ones. A discourse relation can be either **mononuclear** or **multinuclear**. A mononuclear relation connects a nucleus and a satellite (e.g., *Elaboration*, *Attribution* in Figure 1(a)), where a multinuclear relation connects two or more nuclei (e.g., *Joint*, *Contrast*). Thus, an RST discourse tree comprises four types of elements: (i) EDUs that comprise textual information, (ii) the structure or skeleton of the tree, (iii) nuclearity statuses of the discourse units, and (iv) coherence relations by which adjacent discourse units are linked.

Our hypothesis in this article is that the similarity between the discourse trees of an automatic translation and of a reference translation provides additional information that can be valuable for evaluating MT systems. In particular, we believe that better system translations should be more similar to the human translations in their discourse structures than worse ones. As an example, consider the three discourse trees shown in Figure 1: (a) for a reference translation, and (b) and (c) for translations of two different systems from the WMT12 competition. Notice that the tree structure, the nuclearity statuses, and the relation labels in the reference translation are also realized in the system translation in Figure 1(b), but not in Figure 1(c); this makes (b) a better translation compared with (c), according to our hypothesis. We argue that existing metrics that only use lexical and syntactic information cannot distinguish well between the translations in Figure 1(b) and Figure 1(c).

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

### 2.1 Generating Discourse Trees

Conventionally, discourse analysis in RST involves two main subtasks: (*i*) *discourse segmentation*, or breaking the text into a sequence of EDUs, and (*ii*) *discourse parsing*, or the task of linking the discourse units (which could be EDUs or larger units) into labeled discourse trees. Recently, Joty, Carenini, and Ng (2012, 2015) proposed discriminative models for discourse segmentation and discourse parsing. Their discourse segmenter uses a maximum entropy model and achieves state-of-the-art performance with an $F_1$-score of 90.5, whereas human agreement for this task is 98.3 in $F_1$-score.

The discourse parser uses a dynamic Conditional Random Field (Sutton, McCallum, and Rohanimanesh 2007) as a parsing model 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 parse. Using the standard set of 18 coarse-grained discourse relations,[1] the discourse parser achieved an $F_1$-score of 79.8% at the sentence level, which is close to the human agreement of 83%. These high numbers inspired us to develop discourse-aware MT evaluation metrics.[2]

### 2.2 Measuring Similarity Between Discourse Trees

A number of metrics have been proposed to measure the similarity between two labeled trees—for example, Tree Edit Distance (Tai 1979) and various Tree Kernels (TKs) (Collins and Duffy 2001; Smola and Vishwanathan 2003; Moschitti 2006). One advantage of tree kernels is that they provide an effective way to integrate tree structures in kernel-based learning algorithms like SVMs, and learn from arbitrary tree fragments as features.

Collins and Duffy (2001) proposed a syntactic tree kernel to efficiently compute the number of common subtrees in two syntactic trees. To comply with the rules (or productions) of a context-free grammar in syntactic parsing, the subtrees in this kernel are subject to the constraint that their nodes are taken with either all or none of the children. Because the same constraint applies to discourse trees, we use the same tree kernel in our work. Figure 2 shows the valid subtrees according to the syntactic tree kernel for the discourse tree in Figure 1(a). Note that in this work we use the tree kernel only to measure the similarity between two discourse trees rather than to learn subtree features in a supervised kernel-based learning framework like SVM. As an example of the latter, see our more recent work (Guzmán et al. 2014a), which uses tree kernels over syntactic and discourse structures in an SVM preference ranking framework.

Collins and Duffy (2001) proposed two modifications of the kernel when using it in a classifier (e.g., SVM) to avoid the classifier behaving like a nearest neighbor rule: (i) to restrict the tree fragments considered in the kernel computation based on their depth, and/or (ii) to assign relative weights to the tree fragments based on their size. Because we do not use the kernel in a learning algorithm, these modifications do not apply to us; all subtrees are equally weighted in our kernel.

Figure 2 shows that, when applied to discourse trees, the syntactic tree kernel may limit us on the type of substructures that we wish to compare. For example, although matching the complete production (i.e., a parent with all of its children) may make more sense for subtrees with internal nodes only (i.e., non-terminals), we may want to relax this constraint at the terminal (text) level to allow word subsequence matches.

One way to cope with this limitation of the tree kernel is to change the representation of the trees to a form that is suitable to capture the relevant information for our task. For example, in order to allow for the syntactic tree kernel to find subtree matches at the word unigram level, we can include an artificial layer of leaves (e.g., by copying the same *dummy* label below each word). In this way, the words become pre-terminal nodes and can be matched against the words in the other tree.

Apart from this modification to match subtrees at the word level, we experimented with different representations of a discourse tree, each of which produces a different discourse-based evaluation metric. In this section we present two basic representations of the discourse tree, namely, DR and DR-LEX, which we will use in Section 4 to

[1] See Carlson and Marcu (2001) for a detailed description of the discourse relations.

[2] A demo of the parser is available at `http://alt.qcri.org/demos/Discourse_Parser_Demo/`. The source code of the parser is available from `http://alt.qcri.org/tools/discourse-parser/`.

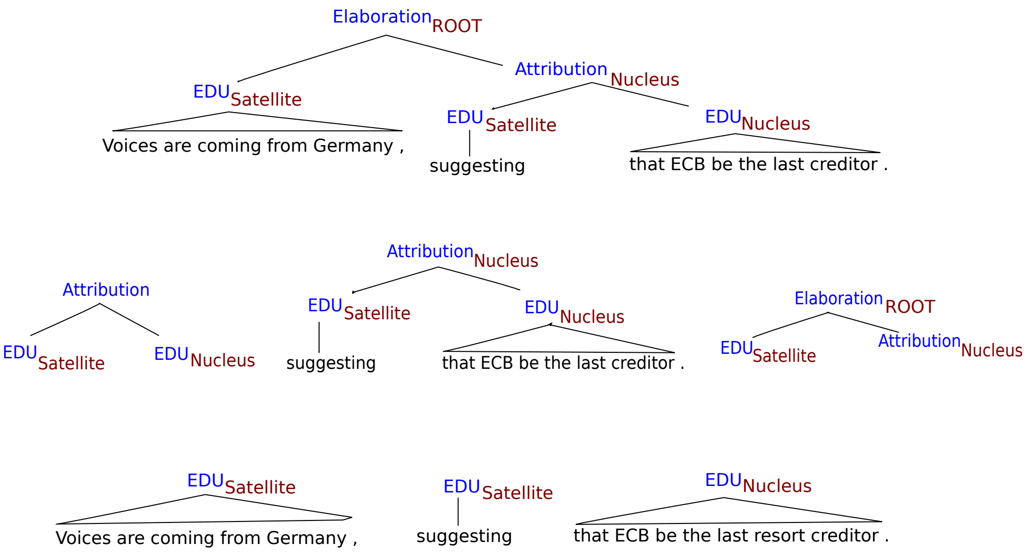
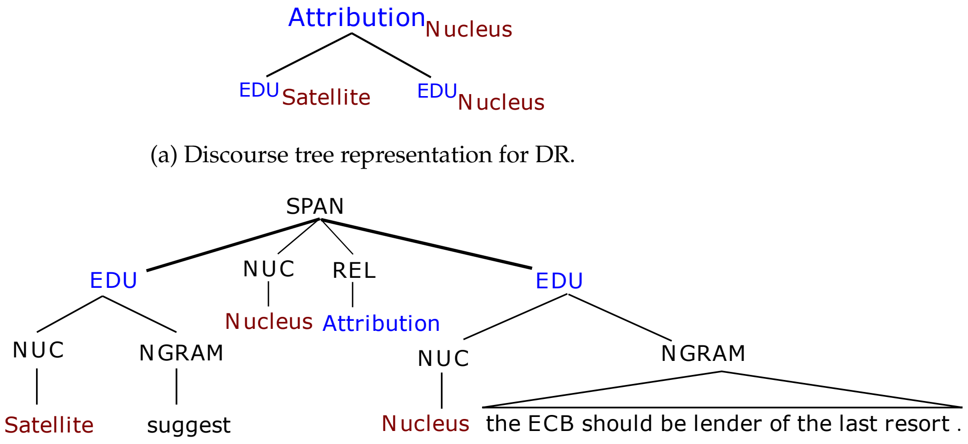


**Figure 2**
Discourse subtrees used by the syntactic tree kernel for the tree in Figure 1(a).

demonstrate that the discourse measures are synergetic with several widely used MT evaluation metrics (DR stands for discourse representation).

Figure 3 shows the two representations DR and DR-LEX for the highlighted subtree in Figure 1b, that spans the text: *suggest the ECB should be the lender of last resort*. As shown in Figure 3(a), DR does not include any lexical item. Therefore, the syntactic tree kernel, when applied to this representation of the discourse tree, measures the similarity between two candidate translations in terms of their discourse representations only.



(a) Discourse tree representation for DR.

(b) Discourse tree representation for DR-LEX

**Figure 3**
Two discourse tree representations for the highlighted subtree in Figure 1(b).

On the contrary, DR-LEX, as shown in Figure 3(b), includes the lexical items to account for lexical matching; moreover, it separates the structure (skeleton) of the tree from its labels (i.e., the nuclearity statuses and the relation labels). This allows the syntactic tree kernel to give partial credit to subtrees that differ in labels but match in their skeletons, or vice versa. More specifically, DR-LEX uses the predefined 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 status and its relation label), whereas an EDU has only one property (its nuclearity status). The words of an EDU are placed under another predefined tag NGRAM. To allow the tree kernel to find subtree matches at the word level, we also include an additional layer of *dummy* leaves (for simplicity, not shown in Figure 3(b)).

## 3. Experimental Setting

In this section, we describe the data sets we used in our experiments, the interpolation approach we applied to combine our discourse-based metrics with pre-existing evaluation metrics, and all the correlation measures we used for evaluation.

### 3.1 Data Sets

In our experiments, we used the data available for the WMT11, WMT12, WMT13, and WMT14 metrics shared tasks for translations into English.[3] This includes the output from the systems that participated in the MT evaluation campaigns in those four years and the corresponding English reference translations. The WMT11 and WMT12 data sets contain 2,000 and 3,003 sentences, respectively, for each of the following four language pairs: Czech–English (CS-EN), French–English (FR-EN), German–English (DE-EN), and Spanish–English (ES-EN). In WMT13, the Russian–English (RU-EN) pair was added to the mix, and the data set has 3,000 sentences for each of the five language pairs. WMT14 dropped ES-EN and included Hindi–English (HI-EN), with each language pair having 3,003 sentences, except for HI-EN, for which there were 2,507 sentences.

The task organizers provided human judgments on the quality of the systems' translations. These judgments represent rankings of the output of five systems chosen at random, for a particular language pair and for a particular sentence. The overall coverage (i.e., the number of unique sentences that were evaluated) was only a fraction of the total (see Table 1). For example, for WMT11 FR-EN, only 247 out of 3,000 sentences have human judgments. Although the evaluation set-up of WMT evaluation is performed in a sentence-level fashion, we believe that it is adequate for our purpose. The annotation interface allowed human judges to take longer-range discourse structure into account, as they were shown the source and the human reference translations in the context of one preceding and one following sentences.[4]

Table 1 shows the main statistics about the data that we used for *training*, where we excluded all pairs for which: (i) both translations were judged as equally good, or (ii) the number of votes for translation$_1$ being better than translation$_2$ equals the number of votes for it being worse than translation$_2$. Moreover, we ignored repetitions—that is, if two judges voted the same way, we did not create two training examples, but just one

[3] `http://www.statmt.org/wmtYY/results.html`, with YY in $\{11, 12, 13, 14\}$.

[4] A detailed description of the WMT evaluation setting can be found in (Bojar et al. 2014).

**Table 1**
Number of systems (sys), unique non-tied translation pairs (pairs), and unique sentences for which such pairs exist (sent) for the different language pairs, for the human evaluation of the WMT11-WMT14 metric shared tasks. These statistics show what we use for *training*; the numbers for *testing* are higher, as explained in the text.

| | **WMT11** | | | **WMT12** | | | **WMT13** | | | **WMT14** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | sys | pairs | sent | sys | pairs | sent | sys | pairs | sent | sys | pairs | sent |
| CS-EN | 12 | 2,477 | 190 | 6 | 8,269 | 937 | 11 | 46,397 | 2,572 | 5 | 10,301 | 1,288 |
| DE-EN | 28 | 7,358 | 346 | 16 | 9,084 | 968 | 17 | 75,856 | 2,589 | 13 | 15,971 | 1,472 |
| ES-EN | 21 | 4,799 | 274 | 12 | 8,751 | 910 | 12 | 36,626 | 2,172 | – | – | – |
| FR-EN | 24 | 5,085 | 247 | 15 | 8,747 | 932 | 13 | 43,234 | 2,272 | 8 | 15,033 | 1,365 |
| RU-EN | – | – | – | – | – | – | 19 | 94,509 | 2,740 | 13 | 24,595 | 1,800 |
| HI-EN | – | – | – | – | – | – | – | – | – | 9 | 14,678 | 1,180 |

(note, however, that on testing, repetitions will be accounted for). Excluding ties and repetitions reduces the number of training pairs significantly (e.g., for WMT13 CS-EN, we have 46,397 pairs, whereas initially there were 85,469 judgments in total).

Note, however, that for *testing*, we used the official full data sets, where we used all pairwise judgments, including judgments saying that both translations are equally good, ties in the number of wins of translation$_1$ vs. translation$_2$ and repetitions. This is important to make our results fully comparable to previously published work.

As a final analysis on the WMT corpora, we studied the complexity of the discourse trees. Recall that we imposed the limitation of working with discourse structures at the sentence level. If we want the discourse metrics to be impactful, we need to make sure that a significant number of sentences have non-trivial discourse trees.

Figure 4 shows the proportion of sentences by discourse tree depth for the WMT11, WMT12, and WMT13 data sets. We computed these statistics with our automatic discourse parser applied to the reference translations. As can be seen, the three data sets show very similar curves. One relevant observation is that more than 70% of the sentences have a non-trivial discourse tree (depth $> 0$). Of course, the proportion of sentences decreases quickly with the tree depth. About 20% of the sentences have trees of depth 2 and slightly over 10% have trees of depth 3. The average depth for the three data sets is 1.77, with a minimum absolute value of 0 and a maximum of 32. The number of EDUs contained in those trees average to 2.77, with a minimum number of 1 and a maximum number of 33. Although the impact of discourse information is potentially higher at the paragraph level or document level, we showed that we have complex enough sentences in our data sets in terms of discourse structure. Thus, we have justified that there is potential in testing the effect of discourse information in MT evaluation metrics.

### 3.2 Learning Interpolation Weights for Metric Combination

In Sections 4.2 and 4.3, we report experiments with a simple linear model combining the predictions of our discourse-based metrics with several pre-existing MT evaluation metrics. The interpolation weights are trained discriminatively from the manually annotated rankings described in the previous section. In order to use the annotations for training, we first transformed the five-way relative rankings into ten pairwise comparisons. For instance, if a judge has ranked the output of systems *A*, *B*, *C*, *D*, *E*

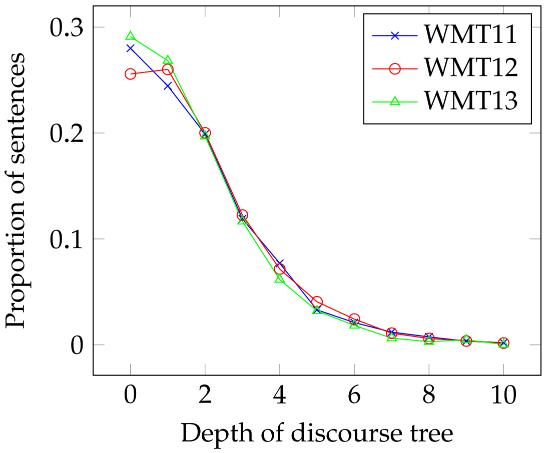


**Figure 4**
Distribution of sentences by tree depth computed based on the reference translations of WMT11, WMT12, and WMT13.

as $A > B > C > D > E$, this would entail the following ten pairwise rankings: $A > B$, $A > C$, $A > D$, $A > E$, $B > C$, $B > D$, $B > E$, $C > D$, $C > E$, and $D > E$. We then use a *maximum entropy* learning framework to learn the interpolation weights, where the classification task is to distinguish a better translation from a worse one for each pair of translation hypotheses. The log likelihood of the training data with $l_2$ regularization on the weight parameters w is:

$$J(\mathrm{w}) = \sum_{i=1}^{N} y_i \log\ \mathrm{Sig}\left(\mathrm{w}^T(\mathrm{u}_\mathrm{i}^1 - \mathrm{u}_\mathrm{i}^2)\right) + (1 - y_i) \log\left(1 - \mathrm{Sig}(\mathrm{w}^T(\mathrm{u}_\mathrm{i}^1 - \mathrm{u}_\mathrm{i}^2))\right) + \lambda \mathrm{w}^T\mathrm{w} \quad (1)$$

where $\mathrm{Sig}(x)$ is the *sigmoid* (aka *logistic*) function, the $\mathrm{u}_\mathrm{i}^1$ and $\mathrm{u}_\mathrm{i}^2$ vectors represent the values of the evaluation measures we are combining for the two translations in the pair $i = (t_1, t_2)$, and $y_i \in \{1, 0\}$ is the human assessment for the pair, that is, $y_i = 1$ if $t_1$ is better than $t_2$, otherwise $y_i = 0$. We learn the model parameters w using the L-BFGS fitting algorithm, which is time- and space-efficient. We learn the regularization strength parameter $\lambda$ using 5-fold cross-validation on the training data set.[5]

Note that our approach to learn the interpolation weights is similar to the one used by PRO for tuning the relative weights of the components of a log-linear SMT model (Hopkins and May 2011). Unlike PRO, (i) we used *human judgments*, not automatic scores, and (ii) we trained on *all pairs*, not on a subsample.

### 3.3 Correlation Measures

In our experiments, we only considered translation into English (as we had a discourse parser for English only), and we used the data described in Table 1. For evaluation, we followed the standard set-up of the Metrics task of WMT12 (Callison-Burch et al. 2012). For segment-level evaluation, we used Kendall's $\tau$ (Kendall 1938), which can be

[5] When fitting the model, we did not include a bias term, as this was harmful.

calculated directly from the human pairwise judgments. For system-level evaluation, we used Spearman's rank correlation (Spearman 1904) and, in some cases, also Pearson correlation (Pearson 1895), which are appropriate correlation measures as here we have vectors of scores.

We measured the correlation of the evaluation metrics with the human judgments provided by the task organizers. As we explained earlier, the judgments represent rankings of the output of five systems chosen at random, for a particular sentence also chosen at random. From each of those rankings, we produce ten pairwise judgments (see Section 3.2). Then, using those pairwise human judgments, we evaluated the performance of the different MT evaluation metrics at the segment or at the system level.

*3.3.1 Segment-Level Evaluation.* We used Kendall's $\tau$ to measure the correlations between the segment-level scores[6] given by a target evaluation metric and the human judgments. Kendall's $\tau$ is defined as follows:

$$\tau = \frac{\#concordant - \#discordant}{\#concordant + \#discordant} \tag{2}$$

where *#concordant* is the number of concordant translation pairs (i.e., pairs for which the human ranking and the corresponding metric scores agree) and *#discordant* is the number of pairs for which the human ranking and the metric score disagree. For example, if the human judgment is that the translation of system $s_i$ for segment $k$ is better than the translation of system $s_j$ for segment $k$, this pair will be considered concordant if the metric gives higher score to $s_i$ than to $s_j$ for segment $k$.

The value of Kendall's $\tau$ ranges between $-1$ (all pairs are discordant) and 1 (all pairs are concordant), and negative values are worse than positive ones. Note that different sets of systems may be ranked for the different segments, but in the calculations we only use pairs of systems for which we have human judgments. Such direct judgments are available for a particular language pair and for a particular segment. We do not calculate Kendall's $\tau$ for each language pair; instead, we consider all pairwise judgments as part of a single set (as implemented in the official WMT scripts).

In the original Kendall's $\tau$ (Kendall 1938), comparisons with human or metric ties are considered neither concordant nor discordant. In the experiments in Section 4, we used the official scorers from the WMT Metrics tasks to compute Kendall's $\tau$. More precisely, in Sections 4.1 and 4.2 we use the WMT12 version of Kendall's $\tau$ (Callison-Burch et al. 2012), whereas in Section 4.3 we report results using the WMT14 scorer (Macháček and Bojar 2014). We used these two different versions of the software to allow a direct comparison to the official results that were reported for the metrics task in WMT12 (Callison-Burch et al. 2012) and WMT14 (Macháček and Bojar 2014).

*3.3.2 System-Level Evaluation.* For the correlation at the system level, we first produce a score for each of the systems according to the quality of their translations based on the evaluation metrics and on the human judgments. Then, we calculate the correlation between the scores for the participating systems using a target metric's scores and the human scores. We do this based on system ranks induced by the scores (using Spearman's rank correlation) or based on the scores themselves (using Pearson correlation). Note that, following WMT, we calculate the correlation score separately

[6] In this section we use the term **segment** instead of **sentence** as we do in the rest of the article, to be consistent with the terminology used in the MT field.

for each language pair, and then we average the resulting correlations to obtain the final score.

*Segment-to-system Score Aggregation.* In order to produce a system-level score based on the pairwise sentence-level human judgments, we need to aggregate these judgments, which we do based on the ratio of wins (ignoring ties), as defined for the official ranking at WMT12 (Callison-Burch et al. 2012):[7]

$$\text{score}(s_i) = \frac{\text{win}(s_i)}{\text{win}(s_i) + \text{loss}(s_i)} \tag{3}$$

where win($s_i$) and loss($s_i$) are the number of wins and losses, respectively, of system $s_i$ against any other system in the segment-level pairwise human judgments.

*Spearman's Rank Correlation.* This is the WMT12 official metric for system-level evaluation. To calculate it, we first convert the raw scores assigned to each system to ranks, and then we use the following formula (Spearman 1904):

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{4}$$

where $d_i$ is the difference between the ranks for system $i$, and $n$ is the number of systems being evaluated. Note that this formula requires that there be no ties in the ranks of the systems (based on the automatic metric or based on the human judgments), which was indeed the case. Spearman's rank correlation ranges between $-1$ and 1. However, unlike Kendall's $\tau$, here the sign does not matter, and high *absolute* values indicate better performance. In our experiments, we used the official script from WMT12.

In some experiments, we also report Pearson correlation (Pearson 1895), which was the official system-level score at WMT14. This is a more general correlation coefficient than Spearman's and does not require that all $n$ ranks be distinct integers. It is defined as follows:

$$r = \frac{\sum_{i=1}^{n} (H_i - \bar{H})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^{n} (H_i - \bar{H})^2}\sqrt{\sum_{i=1}^{n} (M_i - \bar{M})^2}} \tag{5}$$

where $H$ is the vector of the human scores of all participating systems, $M$ is the vector of the corresponding scores as predicted by the given metric, and $\bar{H}$ and $\bar{M}$ are the means for $H$ and $M$, respectively.

The Pearson correlation value ranges between $-1$ and 1, where higher absolute score is better. We used the official WMT14 scoring tool to calculate it.

## 4. Evaluation of the Discourse-Based Metrics

In this section, we show the utility of discourse information for machine translation evaluation. We present the evaluation results at the system level and at the segment level, using our two basic discourse-based metrics, which we refer to as DR and DR-LEX

[7] We use the WMT12 aggregation script. See also Callison-Burch et al. (2012) for a discussion and comparison of several aggregation alternatives.

(Section 2.1). In our experiments, we combine DR and DR-LEX with other evaluation metrics in two different ways: using uniform linear interpolation (at the system level and at the segment level), and using a tuned linear interpolation for the segment-level. We only present the average results over all language pairs. For clarity, in our tables we show results divided into three evaluation groups:

*Group I* contains our discourse-based evaluation metrics, DR, and DR-LEX.

*Group II* includes the publicly available MT evaluation metrics that participated in the WMT12 metrics task, excluding those that did not have results for all language pairs (Callison-Burch et al. 2012). More precisely, they are SPEDE07PP, AMBER, METEOR, TERRORCAT, SIMPBLEU, XENERRCATS, WORDBLOCKEC, BLOCKERRCATS, and POSF.

*Group III* contains other important individual evaluation metrics that are commonly used in MT evaluation: BLEU (Papineni et al. 2002), NIST (Doddington 2002), ROUGE (Lin 2004), and TER (Snover et al. 2006). We calculated the metrics in this group using Asiya. In particular, we used the following Asiya versions of TER and ROUGE: TERP-A and ROUGE-W.[8]

For each metric in groups II and III, we present the system-level and segment-level results for the original metric as well as 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 yield degradation are in *italic*.

For the segment-level evaluation, we further indicate which interpolated results yield statistically significant improvement over the original metric. Note that testing statistical significance is not trivial in our case because we have a complex correlation score for which the assumptions that standard tests make are not met. We thus resorted to a non-parametric randomization framework (Yeh 2000), which is commonly used in NLP research.[9]

### 4.1 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 0.807, which is very close to the BLEU and the TER scores from group II (0.810 and 0.812, respectively). Moreover, DR yields improvements when combined with 13 of the 15 metrics, with a resulting correlation higher than those of the two individual metrics being combined. This fact suggests that DR contains information that is complementary to that used by most of the other metrics.

As expected, DR-LEX performs better than DR because 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, DR-LEX is able to improve 14 out of the 15 metrics. Averaging over all metrics in the table, the combination

[8] These variants are described in page 19 of the ASIYA manual (`http://asiya.lsi.upc.edu/`).

[9] We did not apply the significance test at the system level because of the insufficient number of scores available to sample from; there were a total of 49 (language pair, system) scores for the WMT12 data.

**Table 2**
Results on WMT12 at the system-level (calculated on 6 systems for cs-en, 16 for de-en, 12 for es-en, and 15 for fr-en). Spearman's correlation with human judgments.

| | Metrics | | +DR | +DR-lex |
|---|---|---|---|---|
| **I** | DR | 0.807 | – | – |
| | DR-lex | 0.876 | – | – |
| **II** | SEMPOS | 0.902 | *0.853* | **0.903** |
| | AMBER | 0.857 | *0.829* | **0.869** |
| | Meteor | 0.834 | **0.861** | **0.888** |
| | TerrorCat | 0.831 | **0.854** | **0.889** |
| | SIMPBLEU | 0.823 | **0.826** | **0.859** |
| | TER | 0.812 | **0.836** | **0.848** |
| | BLEU | 0.810 | **0.830** | **0.846** |
| | posF | 0.754 | **0.841** | **0.857** |
| | BlockErrCats | 0.751 | **0.859** | **0.855** |
| | WordBlockEC | 0.738 | **0.822** | **0.843** |
| | XEnErrCats | 0.735 | **0.819** | **0.843** |
| **III** | BLEU | 0.791 | **0.880** | **0.859** |
| | NIST | 0.817 | **0.842** | **0.875** |
| | Rouge | 0.884 | **0.899** | *0.869* |
| | TER | 0.908 | **0.926** | **0.920** |

of DR improves the average of the individual metrics correlation from 0.816 to 0.852 (+0.035) and DR-lex further improves the average results up to 0.868 (+0.052).

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.

### 4.2 Segment-Level Results

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 $-0.433$. This should not be surprising because (i) the discourse tree structure alone does not contain enough information for a good evaluation at the segment level, and (ii) this metric is more sensitive to the quality of the DT, which can be wrong or void. Moreover, 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, allowing us to rank systems more precisely. Because of the low score of DR as an individual metric, it fails to yield improvements when uniformly combined with other metrics (see Untuned +DR column in Table 3).

Again, DR-lex is better than DR; with a positive $\tau$ of 0.133, yet as an individual metric, it ranks poorly compared to other metrics in groups II and III. However, when uniformly combined (see Untuned +DR column) with other metrics, DR-lex outperforms 9 of the 13 metrics in Table 3, with statistically significant improvements in 8 of these cases (p-value <0.01).

**Table 3**
Results on WMT12 at the segment-level (calculated on 11,021 pairs for CS-EN, 11,934 for DE-EN, 9,796 for ES-EN, and 11,594 for FR-EN): untuned and tuned versions. Kendall's $\tau$ with human judgments. Improvements over the baseline are shown in **bold**, and statistically significant improvements are marked with $^{**}$ and $^{*}$ for p-value <0.01 and p-value <0.05, respectively.

| | | | Untuned | | Tuned | |
|---|---|---|---|---|---|---|
| | **Metrics** | **Orig.** | +DR | +DR-LEX | +DR | +DR-LEX |
| **I** | DR | −0.433 | – | – | – | – |
| | DR-LEX | 0.133 | – | – | – | – |
| **II** | SPEDE07PP | 0.254 | *0.190* | *0.223* | *0.253* | 0.254 |
| | METEOR | 0.247 | *0.178* | *0.217* | **0.250** | **0.251** |
| | AMBER | 0.229 | *0.180* | *0.216* | **0.230** | **0.232** |
| | SIMPBLEU | 0.172 | *0.141* | **0.191**$^{**}$ | **0.181**$^{**}$ | **0.199**$^{**}$ |
| | XENERRCATS | 0.165 | *0.132* | **0.185**$^{**}$ | **0.175**$^{**}$ | **0.194**$^{**}$ |
| | POSF | 0.154 | *0.125* | **0.201**$^{**}$ | **0.160**$^{**}$ | **0.201**$^{**}$ |
| | WORDBLOCKEC | 0.153 | *0.122* | **0.181**$^{**}$ | **0.161**$^{**}$ | **0.189**$^{**}$ |
| | BLOCKERRCATS | 0.074 | *0.068* | **0.151**$^{**}$ | **0.087**$^{**}$ | **0.150**$^{**}$ |
| | TERRORCAT | −0.186 | **−0.111** | **−0.104**$^{**}$ | **0.181**$^{**}$ | **0.196**$^{**}$ |
| **III** | BLEU | 0.185 | *0.154* | **0.190** | **0.189** | **0.194**$^{*}$ |
| | NIST | 0.214 | *0.172* | *0.206* | **0.222**$^{**}$ | **0.224**$^{**}$ |
| | ROUGE | 0.185 | *0.144* | **0.201**$^{**}$ | **0.196**$^{**}$ | **0.218**$^{**}$ |
| | TER | 0.217 | *0.179* | **0.229**$^{**}$ | **0.229**$^{**}$ | **0.246**$^{**}$ |

Following the learning method described in Section 3.2, we experimented also with tuning the interpolation weights in the metric combinations. We report results for (i) cross-validation on WMT12, and (ii) tuning on WMT12 and testing on WMT11.

*Cross-validation on WMT12.* For cross-validation on 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 then used the remaining fold for evaluation. Note that in this process, we aggregated the data for different language pairs, and we produced a single set of tuning weights for all language pairs.[10]

The results are shown in the last two columns of Table 3 (Tuned). We can see that the tuned combinations with DR-LEX improve over all but one of the individual metrics in groups II and III, with statistically significant differences in 10 out of the 12 cases. Even more interestingly, the tuned combinations that include the much weaker metric DR now improve over 12 out of 13 of the individual metrics, with 9 of these differences being statistically significant with p-value <0.01. This is remarkable given that DR has a strong negative $\tau$ as an individual metric at the sentence-level. Again, these results suggest that both DR and DR-LEX contain information that is complementary to that of the individual metrics that we experimented with.

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

Averaging over all 13 cases, DR improves Kendall's $\tau$ from 0.159 to 0.193 (+0.035), and DR-LEX improves it to 0.211 (+0.053). These sizable improvements highlight the importance of tuning the linear combination when working at the segment level.

*Testing on WMT11.* To rule out the possibility that the improvement of the tuned metrics on WMT12 could have come from over-fitting, and also in order to verify that the tuned metrics generalize when applied to other sentences, we also tested on an additional data set: WMT11. We tuned the weights for our metric combinations on *all* WMT12 pairwise judgments (no cross-validation), and we evaluated them on the WMT11 data set. Because 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), this time we only report results for the standard group III metrics, thus ensuring that the metrics in the experiments are consistent for 2011 and 2012.

The results, presented in Table 4, show the same pattern as before: (i) adding DR and DR-LEX improve overall individual metrics, with the differences being statistically significant in seven out of the eight cases with p-value <0.01; and (ii) the contribution of DR-LEX is consistently larger than that of DR. Observe that 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.

### 4.3 DR-Based Metrics in a Strong MT Evaluation Measure

From the results presented in the previous sections, we can conclude that discourse structure is an important information source, which is not entirely correlated to other information sources considered so far, and thus should be taken into account when designing future metrics for automatic evaluation of machine translation output. In this section we show how the simple combination of DR-based metrics with a selection of other existing strong MT evaluation metrics can lead to a very competitive evaluation metric, DISCOTK$_{party}$ (Joty et al. 2014), which we presented at the metrics task of WMT14 (Macháček and Bojar 2014).

ASIYA (Giménez and Màrquez 2010a) is a suite for MT evaluation that provides a large set of metrics using different levels of linguistic information. We used the 12 individual metrics from ASIYA's ULC (Giménez and Màrquez 2010b), which was

**Table 4**
Results on WMT11 at the segment-level (calculated on 3,695 pairs for CS-EN, 8,950 for DE-EN, 5,974 for ES-EN, and 6,337 for FR-EN): tuning on the entire WMT12. Kendall's $\tau$ with human judgments. Improvements over the baseline are shown in **bold**, and statistically significant improvements are marked with ** for p-value <0.01.

| | | | Tuned | |
|---|---|---|---|---|
| | **Metrics** | **Orig.** | +DR | +DR-LEX |
| **I** | DR | −0.447 | – | – |
| | DR-LEX | 0.146 | – | – |
| **III** | BLEU | 0.186 | **0.192** | **0.207**** |
| | NIST | 0.219 | **0.226**** | **0.232**** |
| | ROUGE | 0.205 | **0.218**** | **0.242**** |
| | TER | 0.262 | **0.274**** | **0.296**** |

**Table 5**
Comparing our tuned metric to the best rivaling metric at WMT14, for each individual language pair (this best rival differs across language pairs) at the segment-level using Kendall's τ. Statistically significant improvements are marked with ** for p-value < 0.01.

| **System** | FR-EN | DE-EN | HI-EN | CS-EN | RU-EN | Overall |
|---|---|---|---|---|---|---|
| DISCOTK$_{party}$ | **0.433**** | **0.380**** | 0.434 | **0.328**** | **0.355**** | **0.386**** |
| Best at WMT14 | 0.417 | 0.345 | **0.438** | 0.284 | 0.336 | 0.364 |
| | +0.016 | +0.035 | −0.004 | +0.044 | +0.019 | +0.024 |

the best performing metric both at the system level and at the segment level at the WMT08 and WMT09 metrics tasks. From the original ULC, we replaced METEOR by the four newer variants METEOR-ex (exact match), METEOR-st (+stemming), METEOR-sy (+synonymy lookup), and METEOR-pa (+paraphrasing) in ASIYA's terminology (Denkowski and Lavie 2011). We also added to the mix TERp-A (a variant of TER with paraphrasing), BLEU, NIST, and ROUGE-W, for a total of 18 individual metrics. The metrics in this set use diverse linguistic information, including lexical-, syntactic-, and semantic-oriented individual metrics.

Regarding the discourse metrics, we used five variants, including DR and DR-LEX described in Section 2, and three more constrained variants oriented to match words between trees only if they occur under the same substructure types (e.g., the same nuclearity type). These variants are designed by introducing structural modifications in the discourse trees. A detailed description can be found in Joty et al. (2014).

We tuned the relative weights of the previous 23 individual metrics (18 ASIYA+ 5 discourse) following the same maximum entropy learning framework described in Section 3.2. As the training set, we used the simple concatenation of WMT11, WMT12, and WMT13.

DISCOTK$_{party}$ was the best-performing metric at WMT14 both at the segment and at the system level, among a set of 16 and 20 participants, respectively (Macháček and Bojar 2014). Table 5 shows a comparison at the segment level of our tuned metric DISCOTK$_{party}$ to the best rivaling metric at WMT14, for each individual language pair, using Kendall's τ. Note that this best rival differs across language pairs, for example, for FR-EN, HI-EN, and CS-EN it is BEER, for DE-EN it is UPC-STOUT, and for RU-EN it is REDCOMBSENT. We can see that our metric outperforms this best rival for four of the language pairs, with statistically significant differences. The only exception is HI-EN, where the best rival performs slightly better, not statistically significantly.

System translations for Hindi–English were of extremely low quality, and were very hard to discourse-parse accurately.[11] The linguistically heavy components of our DISCOTK$_{party}$ (discourse parsing, syntactic parsing, semantic role labeling, etc.) may suffer from the common ungrammaticality of the translation hypotheses for HI-EN, whereas other, less linguistically heavy metrics seem to be more robust in such cases.

We show in Figure 5 the weights for the individual metrics combined in DISCOTK$_{party}$ after tuning on the combined WMT11+12+13 data set. The horizontal axis displays all the individual metrics involved in the combination. The first block of metrics (from BLEU to DR-Orp*) consists of the 18 ASIYA metrics. The last five (from

[11] Note that the Hindi–English language pair caused a similar problem for a number of other metrics at the WMT14 Shared Task competition that relied on linguistic analysis.

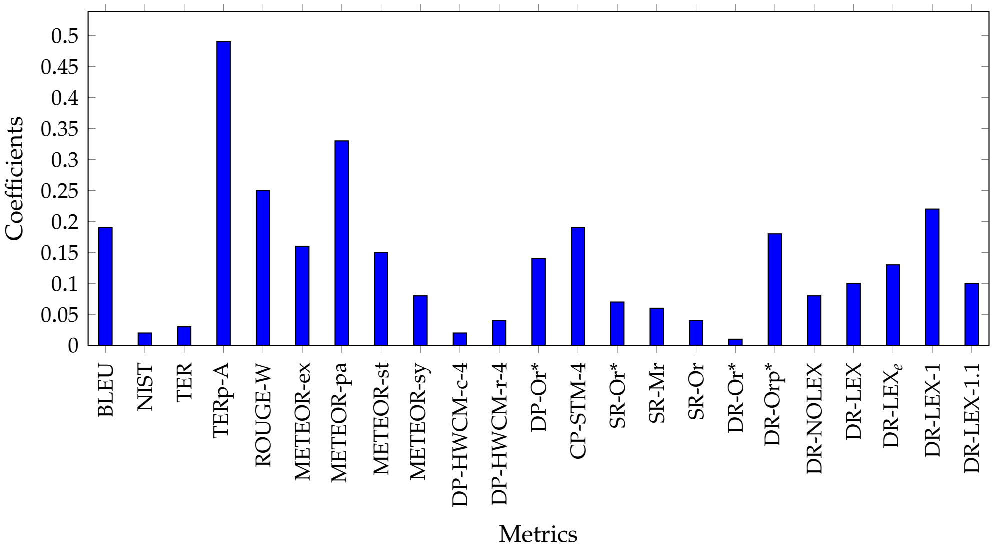


**Figure 5**
Absolute coefficient values after tuning the DiscoTK$_{party}$ metric on the WMT11+12+13 data set.

DR-NOLEX to DR-LEX$_{1.1}$) metrics are the metric variants based on discourse trees. Note that all metric scores are passed through a *min-max* normalization step to put them in the same scale before tuning their relative weights.

We can see that most of the metrics involved in the metric combination play a significant role, the most important ones being TERp-A, METEOR-pa (paraphrases), and ROUGE-W. Some metrics accounting for syntactic and semantic information also get assigned relatively high weights (DP-Or*, CP-STM-4, and DR-Orp*). Interestingly, all five variants of our discourse metric received moderately high weights, with the four variants using lexical information (DR-LEX's) being more important. In particular, DR-LEX$_1$ has the fourth highest absolute weight in the overall combination. This confirms again the importance of discourse information in machine translation evaluation.

## 5. Analysis

When dealing with evaluation metrics based on lexical matching, such as BLEU or NIST, it is easier to understand how and why they work, and what their limitations are. However, if a metric deals with complex structures like discourse trees, it is not straightforward to explain its performance.

In this section, we aim to better understand which parts of the discourse trees have the biggest impact on the performance of the discourse-based measures presented in Section  2. For that purpose, we first conduct an ablation study (see Section 5.1), where we dissect the different components of the discourse trees, and we analyze the impact that the deletion of such components has on the performance of our evaluation metrics. In a second study (see Section 5.2), we analyze which parts of a complete discourse tree are most useful to distinguish between good and bad translations. Overall, the components that we focus on in our analysis are the following: (i) Discourse relations (Elaboration, Attribution, etc.); (ii) Nuclearity statuses (i.e., Nucleus and Satellite); and

(iii) Discourse structure (boundaries of the elementary discourse units, depth of the tree, etc.).

The previous two studies focus on quantitative aspects of the discourse trees. Section 5.3 discusses one real example to understand from a more qualitative point of view the contribution of the sentence-level discourse trees in the evaluation of good and bad translations. Finally, in Section 5.4, we discuss the issue of whether discourse trees provide information that is complementary to syntax.

### 5.1 Ablation Study at the System Level

We analyze the performance of our discourse-based metric DR-LEX at the system level. We use DR-LEX instead of DR, as it exhibits the most competitive performance, and incorporates both lexical and discourse information. We selected system-level evaluation because the metric is much more stable and accurate at the system level than at the segment level.

In our ablation experiments we contrast the original DR-LEX metric, computed over full RST trees, to variations of the same, where the discourse trees have less information. When removing a particular element, we replace the corresponding labels by a *dummy* tag ($*$). We have the following ablation conditions, which are illustrated in Figure 6:

1. *Full*: Original DR-LEX metric with the full (labeled) RST tree structure.
2. *No discourse relations*: We replace all relation labels (Attribution, Elaboration, etc.) in the tree by a *dummy* tag.
3. *No nuclearity*: We replace all the nuclearity statuses (i.e., Nucleus, Satellite), by a *dummy* tag.
4. *No relation and no nuclearity tags*: We replace both the relation and the nuclearity labels by dummy tags. This leaves the discourse structure (i.e., the skeleton of the tree) along with the lexical items.
5. *No discourse structure*: We remove all the discourse structure, and we only leave the lexical information. Under this representation, the evaluation metric corresponds to unigram lexical matching.

We scored all modified trees using the same tree kernel that we used in DR-LEX, and we scored their resulting rankings accordingly. The summarized system-level results for WMT11-13 are shown in Table 6, where we used all into-English language pairs.

We can see a clear pattern in Table 6. Starting from the lexical matching (*no_discourse*), each layer of discourse information helps to improve performance, even if just a little bit. Overall, we observe a cumulative gain from 0.849 to 0.881 in terms of Spearman's ρ. Having only the discourse structure (*no_nuc & no_rel*) improves the performance over using lexical items only. This means that identifying the boundaries of the discourse units in the translations (i.e., which lexical items correspond to which EDU), and how those units should be linked, already can tell us something about the quality of the translation. Next, by adding nuclearity information (*no_rel*), we observe further improvement. This means that knowing which discourse unit is the main one and which one is subordinate is helpful for assessing the quality of the translation. Finally, using the discourse structure, the nuclearity, and the relations together yields the best overall performance. The differences are not very large, but the tendency is consistent across data sets.

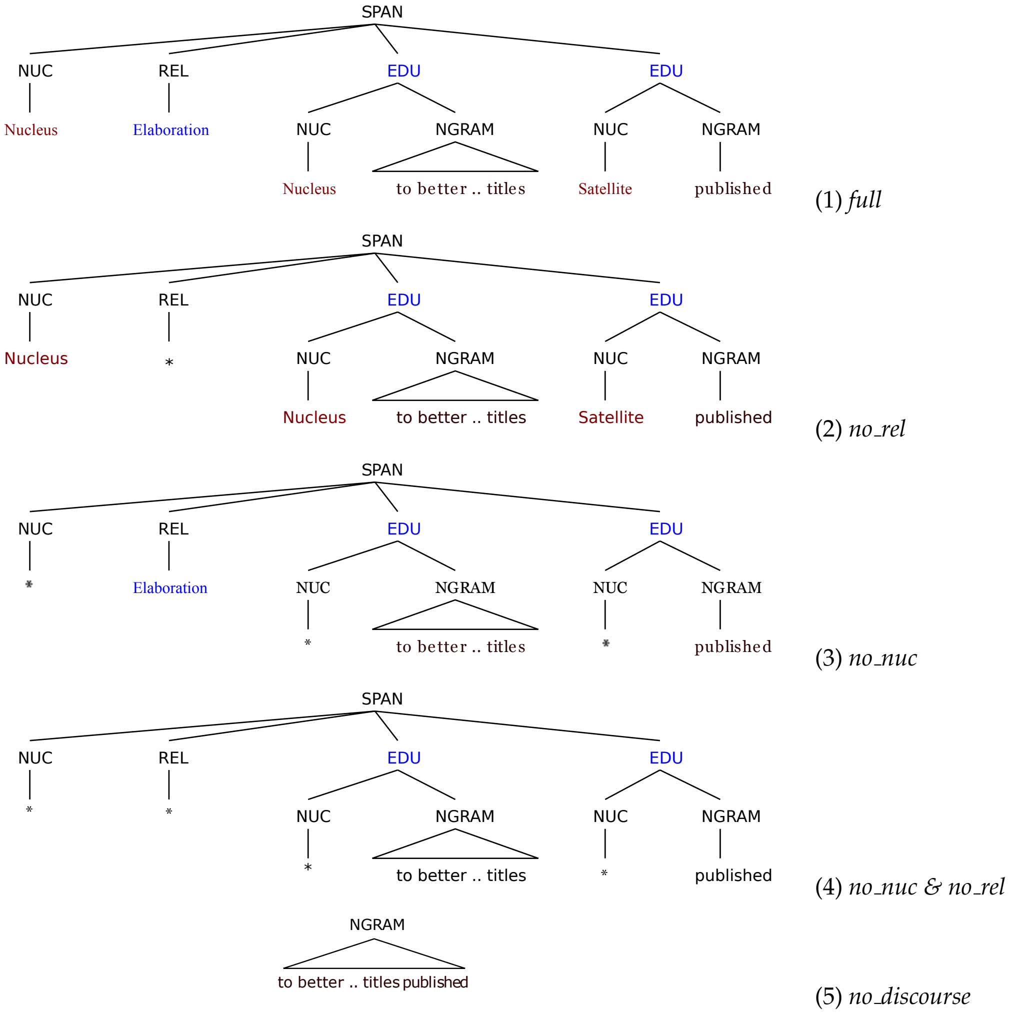


**Figure 6**
Different discourse trees for the same example translation: "to better link the creation of content for all the titles published," with decreasing amount of discourse information. The five representations correspond to the ones used in the ablation study.

**Table 6**
System-level Spearman (ρ) and Pearson (*r*) correlation results for the ablation study over the DR-LEX metric across the WMT{11,12,13} data sets and overall.

| | | 2011 | | 2012 | | 2013 | | Overall | |
|---|---|---|---|---|---|---|---|---|---|
| | **RST variant** | ρ | *r* | ρ | *r* | ρ | *r* | ρ | *r* |
| **DR-LEX** | *full* | **0.848** | **0.860** | **0.876** | **0.912** | **0.920** | **0.919** | **0.881** | **0.897** |
| | *no_rel* | 0.843 | 0.856 | 0.876 | 0.909 | 0.919 | 0.919 | 0.879 | 0.895 |
| | *no_nuc* | 0.822 | 0.828 | 0.867 | 0.896 | 0.910 | 0.914 | 0.866 | 0.879 |
| | *no_nuc & no_rel* | 0.815 | 0.826 | 0.847 | 0.891 | 0.915 | 0.913 | 0.859 | 0.877 |
| | *no_discourse* | 0.794 | 0.798 | 0.865 | 0.863 | 0.887 | 0.903 | 0.849 | 0.855 |

**Table 7**
System-level Spearman (ρ) and Pearson (*r*) correlation results for the ablation study over the DR-LEX metric across language pairs for the WMT{11,12,13} data sets.

| | | CS-EN | | DE-EN | | ES-EN | | FR-EN | |
|---|---|---|---|---|---|---|---|---|---|
| | **RST variant** | ρ* | *r* | ρ | *r* | ρ | *r* | ρ | *r* |
| **DR-LEX** | *full* | 0.890 | **0.893** | **0.782** | **0.840** | **0.970** | **0.952** | **0.943** | **0.940** |
| | *no_rel* | 0.894 | 0.892 | 0.775 | 0.833 | 0.968 | 0.952 | 0.942 | 0.939 |
| | *no_nuc* | **0.899** | 0.885 | 0.739 | 0.802 | 0.958 | 0.949 | 0.935 | 0.925 |
| | *no_nuc & no_rel* | 0.895 | 0.884 | 0.720 | 0.798 | 0.935 | 0.950 | 0.940 | 0.917 |
| | *no_discourse* | 0.833 | 0.861 | 0.738 | 0.743 | 0.942 | 0.930 | 0.936 | 0.919 |

Interestingly, the nuclearity status (*no_rel*) is more important than the type of relation (*no_nuc*). Eliminating the latter yields a tiny decrease in performance, whereas ignoring the former causes a much larger drop. Although this might seem counterintuitive at first (because we think that knowing the type of discourse relation is important), this can be attributed to the difficulty of discourse parsing machine translated text. As we will observe in the next section, assigning the correct relation can be a much harder problem than predicting the nuclearity statuses. Thus, parsing errors might be undermining the effectiveness of the discourse relation information.

Table 7 presents the results of the same ablation study but this time broken down per language pair. For each language pair, all years are considered (2011–2013). Overall, we observe the same pattern as in Table 6, namely, that all layers of discourse information are helpful to improve the results, and that the nuclearity information is more important than the discourse relation types.[12]

However, some differences are observed depending on the language pair. For example, Spanish–English exhibits larger improvements (ρ goes from 0.942 to 0.970) than French-English (ρ goes from 0.936 to 0.943). This is despite both language pairs being mature in terms of the expected quality for these systems. On another axis, the German–English language pair shows much lower overall correlation compared to Spanish–English (0.782 vs. 0.970). This can be the effect of the inherent difficulty of this language pair because of long-distance reordering and so forth. However, note that adding all the discourse layers increases ρ from 0.738 to 0.782. These observations are consistent with our findings in the next section, where we explore the different parts of the discourse trees at a more fine-grained level.

## 5.2 Discriminating Between Good and Bad Translations

In the previous section, we analyzed how different parts of the discourse tree contribute to the performance of the DR-LEX metric. In this section, we take a different approach: We investigate whether the information contained in the discourse trees helps to differentiate good from bad translations.

In order to do so, we analyze the discourse trees generated for three groups of translations: (i) *gold*, the reference translations; (ii) *good*, the translations of the top-two best (per language pair) systems; and (iii) *bad*, the translations of the worst-two

[12] Note that the results of Spearman's ρ for CS-EN do not follow exactly the same pattern. This instability might be due to the small number of systems for this language pair (see Table 1).

(per language pair) systems. Our hypothesis is that there are characteristics in the *good*-translation discourse trees that make them more similar to the *gold*-translation trees than the *bad*-translation trees. The characteristics we analyze here are the following: relation labels, nuclearity labels, tree depth, and number of words. We perform the analysis at the sentence level, by comparing the trees of the *gold*, *good*, and *bad* translations.

*5.2.1 Discourse Relations.* There are 18 discourse relation labels in our RST parser. We separately computed the label frequency distributions from the RST trees of all *gold*, *good*, and *bad* translation hypotheses. Figure 7 shows the histogram for the ten most frequent classes on the Spanish–English portion of the WMT12 data set. We can see that there are clear differences between the *good* and the *bad* distributions, especially in the frequencies of the most common tags (*Elaboration* and *Same-Unit*). The *good* hypotheses have a distribution that is much closer to the human references (*gold*). For example, the frequency difference for the *Elaboration* tag between *good* and *gold* translation trees is 58, which is smaller than the difference between *bad* and *gold*, 323. In other words, the trees for *bad* translations exhibit a surplus of *Elaboration* tags.

If we compare the entire frequency distribution across different relations for the whole WMT12, we observe that the Kullback–Leibler (KL) divergence (Kullback and Leibler 1951) between the *good* and the *gold* distributions is also smaller than the KL divergence between the *bad* and *gold*: 0.0021 vs 0.0039, and a similar tendency holds for WMT13. This means that *good* translations have discourse trees that encode relation tags that match the *gold* translation trees. This suggests that the relation tags should be an important part of the discourse metric.

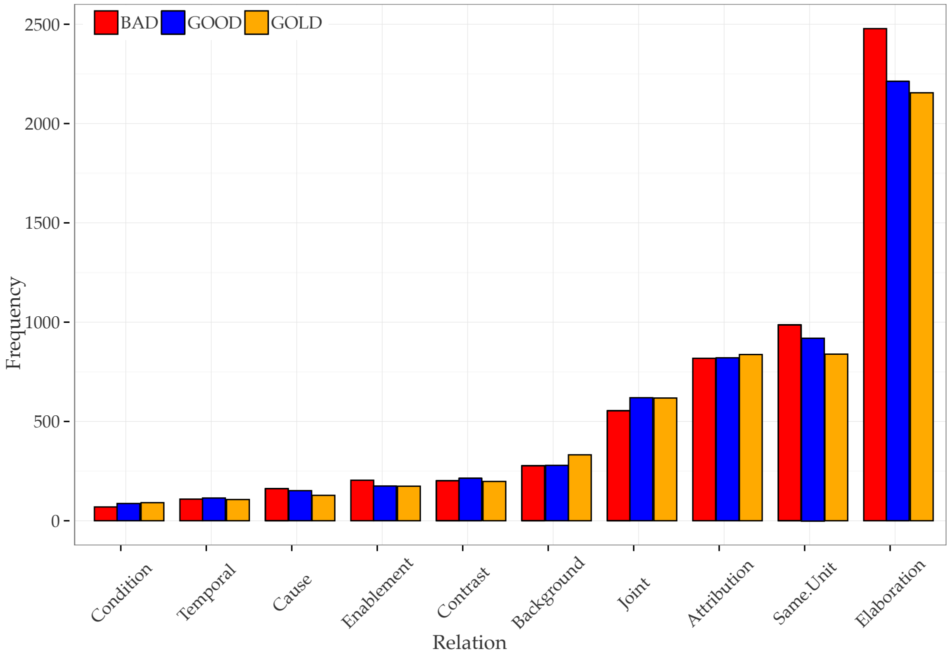


**Figure 7**
Distribution of discourse relations for *gold*, *good*, and *bad* automatic translations on WMT12 Spanish–English. We show the ten most frequent discourse relations only.

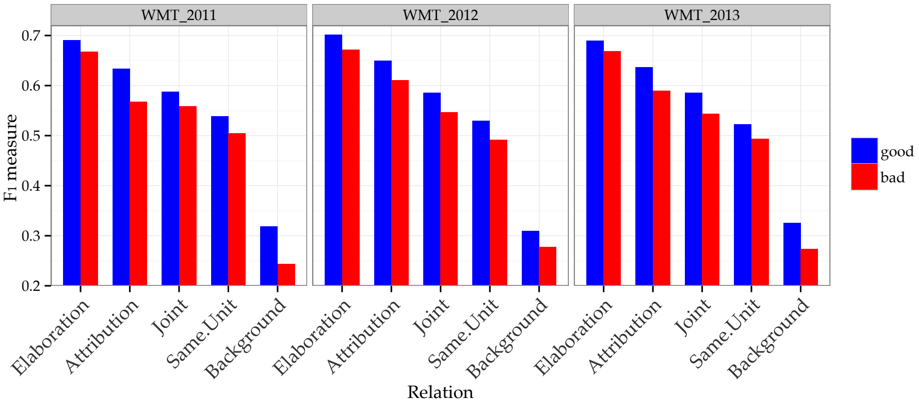


**Figure 8**
$F_1$ score for each of the top-five relations in *good*- vs. *bad*-translation trees across the WMT{11,12,13} data sets.

In a second step, we computed the micro-averaged $F_1$ score for each relation label, taking the *gold* translation discourse trees as a reference. Note that computing standard parsing quality metrics that span over constituents (e.g., $F_1$ score over the constituents), would require the leaves of the two trees to be the same. In our case, we work with two different translations (one *gold* and one MT-generated), which makes their RST trees not directly comparable. Therefore, we apply an approximation, and we measure $F_1$ score over the total number of instances of a specific tag, regardless of their position in the tree. Furthermore, we also consider that every instance of a predicted tag is correct if there is a corresponding tag of the same type in the *gold* tree. Effectively, this makes the number of *true positives* for a specific tag equal to the minimum number of instances for that tag in either the hypothesis or the *gold* trees. Although this is a simplification, this gives us an idea of how closely the RST trees for *good*/*bad* translation approximate the trees from the references.

The results for the five most prevalent relations are shown in Figure 8. We can see systematically higher $F_1$ scores for *good*-translation trees compared with *bad* ones across all relations and all corpora. This supports our hypothesis at the discourse relation level—that is, discourse trees for *good* translations contain more similar discourse labels to the reference translation trees. Note, however, that $F_1$ scores vary across relations and they are not very high (highest is around 70%), indicating that they are hard to predict.

Figure 9 contains the same information, but this time broken down by language pair. For each language pair and corpus year, we micro-average the results for the five most frequent discourse relations. Again, we observe a clear advantage for the *good*-translation trees over the *bad* ones for all language pairs and for all years. Some differences are observed across language pairs, which do not always have an intuitive explanation in terms of the difficulty of the language pair.[13] For instance, larger gaps are observed for ES-EN and DE-EN, compared to the rest. This correlates very well with

[13] Differences between language pairs can be attributable to the particular MT systems that participated in the competition, which is a variable that we cannot control for in these experiments.

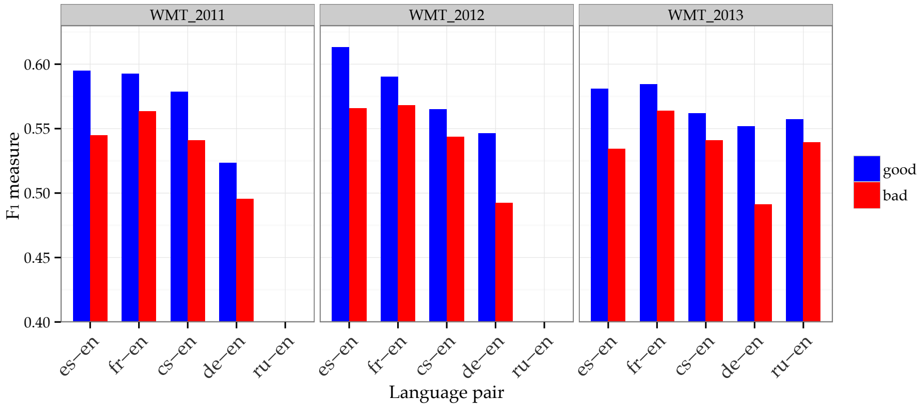


**Figure 9**
$F_1$ score for each language pair in *good*- vs. *bad*-translation trees across the WMT{11,12,13} data sets: micro-averaging the scores of the top-five relations.

the results in Table 7, clearly connecting the discourse similarity and the quality of the evaluation metrics.

*5.2.2 Nuclearity and Other Tree Information.* **Nuclearity** describes the role of the discourse unit within the relation, which can be central (*Nucleus*) or supportive (*Satellite*). Here, we study the distribution of these labels together with two extra elements from the trees: the EDUs and the depth of the discourse tree (Depth). The results are shown in Figure 10. For the number of *Nucleus*, *Satellite*, and EDU labels, we compute the simplified $F_1$ scores in the same way that we did for relation labels, focusing on the number of instances. For the tree Depth, we compute the micro-averaged root-mean-squared-error, or RMSE.

As with the discourse relations, we observe better results for the nuclearity labels and the other tree elements from the *good*-translation trees, compared with the *bad* ones. This is consistent across all years (higher $F_1$ or lower RMSE). Note that the $F_1$ values for nuclearity labels are significantly higher than the $F_1$ scores for discourse relations (now moving in the 0.78–0.82 interval, compared with $F_1$ average scores below 0.60 in the case of discourse relations). This helps to explain the larger impact of the nuclearity elements in the evaluation measure (see Tables 6 and 7). Since predicting discourse segments is easier than predicting nuclearity labels ($F_1$ values close to 0.89), the EDU structure contributes to improving the evaluation measure; this corresponds mainly to the *no_nuc & no_rel* case in the ablation study (again, see Tables 6 and 7).

Finally, Figure 11 shows the results by language pair. We show the micro-averaged $F_1$ scores of the nuclearity labels and EDUs (upper charts), and the RMSE for Depth (lower charts). Once again, the $F_1$ and RMSE results for *good* translations are better than those for *bad* ones, sometimes by large margins. The only exception is for Depth in FR-EN (WMT13). Looking at the overall scores and at the size of the gaps between the scores for *good* and *bad*, we can see that they are consistent with the per-language results of Table 7, showing once again the direct relation between matching discourse elements and the correlation with the human assessments of the discourse-based DR-LEX metric.

The main conclusions that we can draw from this analysis can be summarized as follows: (i) The similarity between discourse trees is a good predictor of the quality of

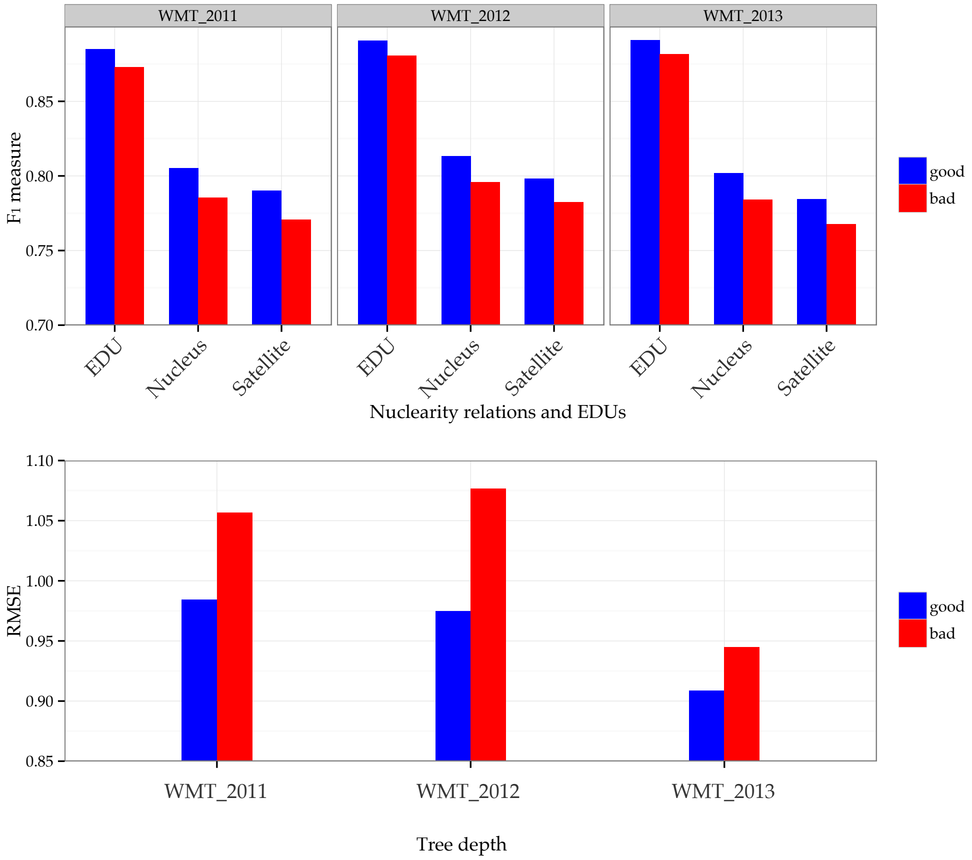


**Figure 10**
$F_1$ scores for the nuclearity relations and EDUs (upper chart), and RMSE for Depth (lower chart) in *good*- vs. *bad*-translation trees, across the WMT{11,12,13} data sets.

the translation, according to the human assessments; (ii) Different levels of discourse structure and relations provide different information, which shows smooth accumulative contribution to the final correlation score; (iii) Both discourse relations and nuclearity labels have sizeable impact on the evaluation metric, the latter being more important than the former. The last point emphasizes the appropriateness of the RST theory as a formalism for the discourse structure of texts. Contrary to other discourse theories (e.g., the Discourse Lexicalized Tree Adjoining Grammar [Webber 2004] used to build the Penn Discourse Treebank [Prasad et al. 2008]), RST accounts for the nuclearity as an important element of the discourse structure.

### 5.3 Qualitative Analysis of Good and Bad Translations

In the previous two sections we provided a quantitative analysis of which discourse information has the biggest impact on the performance of our discourse-based measure (DR-LEX) and also which parts of the discourse trees help in distinguishing *good* from *bad* translations. In this section, we present some qualitative analysis by inspecting a

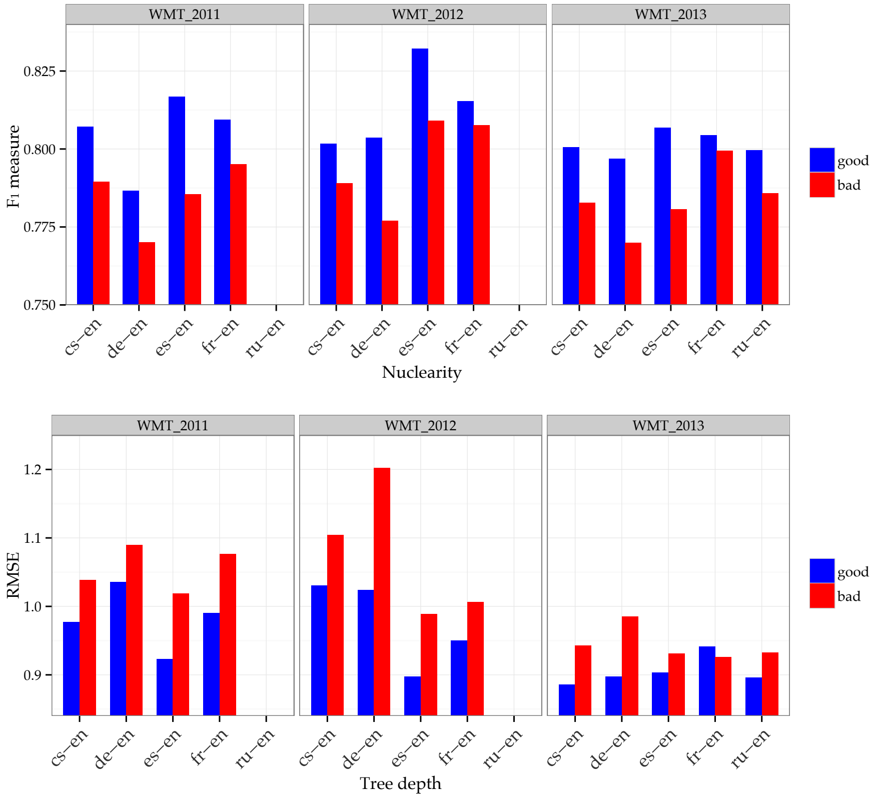


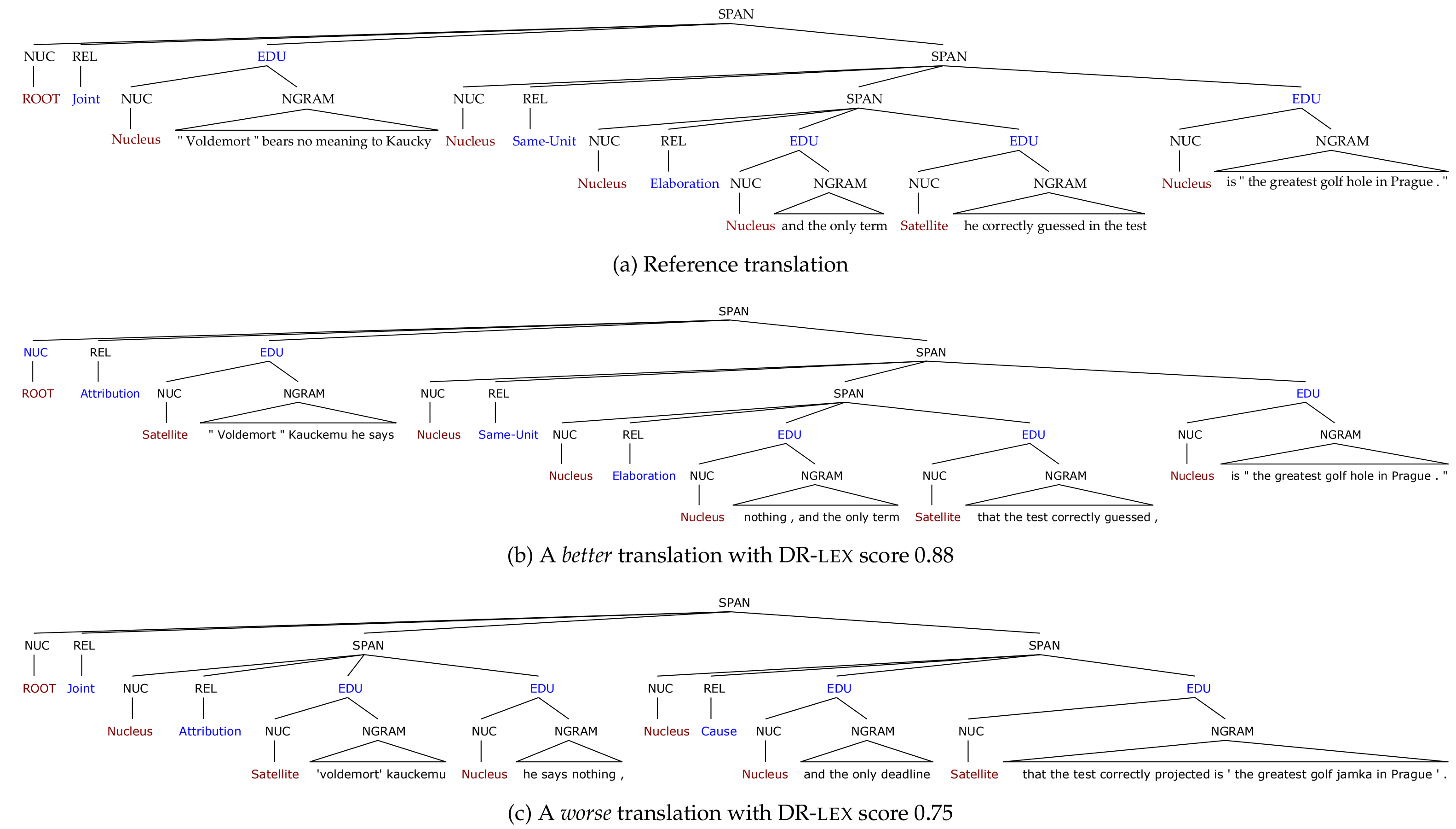
**Figure 11**
Micro-averaged $F_1$ scores for each language pair for nuclearity and EDU, and RMSE for Depth in *good*- vs. *bad*-translation trees across the WMT{11,12,13} data sets.

real example of *good* vs. *bad* translations, and showing how the discourse trees help in assigning similarity scores to distinguish them.

Figure 12 shows a real example with discourse trees for a reference (a) and two alternative translations, one (b) being better than the second (c). The examples are extracted from the WMT11 data set (CS-EN), and the discourse trees are obtained with our automatic discourse parser. Discourse trees are presented with the unfolded format introduced in Figure 3(b).

Translation 12(b) gets a DR-LEX score of 0.88, which is higher than the score for translation 12(c), 0.75. Part of the difference is explained by the fact that translation 12(b) provides better word-based translation, including complete EDU constituents (e.g., *is "the greatest golf hole in Prague"*). But also, 12(b) obtains many more subtree matches with the reference at the level of the discourse structure. This translation has the same discourse structure and labels as the reference, with the only exception of the top-most discourse relation (*Joint* vs. *Attribution*). This tendency is observed across the data sets, and it is quantitatively verified in previous Section 5.2 (i.e., *good* translations tend to share the tree structure and labels with the reference translations).

On the other hand, translation 12(c) is much more ungrammatical. This leads to inaccurate parsing, producing a discourse tree that is flatter than the reference discourse



(a) Reference translation

(b) A *better* translation with DR-LEX score 0.88

(c) A *worse* translation with DR-LEX score 0.75

**Figure 12**
Example of discourse trees for good and bad translations in comparison with the reference translation. Example extracted from WMT-2011 (CS-EN).

tree and that has many more inaccuracies at the discourse relation level. Consequently, the tree kernel finds fewer subtree matches, and the similarity score becomes lower.

Note that the proposed kernel-based similarity assigns the same weight to all subtree matches encountered, so it is not possible for the metric to modulate which are the most important features to distinguish better from worse translations.[14] The success of the metric is based solely on the assumption (verified in Section 5.2) that better translations will exhibit discourse trees that are closer to the reference. A natural step to follow would be to try to learn with preference and convolutional kernels which of these subtree structures (understood as implicit features) help to discriminate better from worse translations. This is the approach followed by Guzmán et al. (2014a), which is also mentioned in Section 5.4.

## 5.4 Does Discourse Provide Relevant Information Beyond Syntax?

Discourse parsing at the sentence level relies heavily on syntactic features extracted from the syntactic parse tree. One valid question to raise is whether the sentence-level discourse structure provides any relevant information for MT evaluation apart from the syntactic relations. Note that in Section 4.3 we combined up to 18 metrics with our discourse-based evaluation metrics. Three of them use dependency parsing features (DP-HWCM-c-4, DP-HWCM-r-4, and DP-Or*; cf. Figure 5) and a fourth one uses constituency parse trees (CP-STM-4; Figure 5).[15] According to the interpolation weights, the contribution of these metrics is not negligible, but it seems to be lower than that of the DR metrics. Still, this is a too indirect way of approaching the comparison.

Our previous work (Guzmán et al. 2014a) helps answer the question about the complementarity of the two sources of information in a more direct way. In that paper, we proposed a pairwise setting for learning MT evaluation metrics with preference tree kernels. The setting can incorporate syntactic and discourse information encapsulated in tree-based structures and the objective is to learn to differentiate better from worse translations by using all subtree structures as implicit features. The discourse parser we used is the same used in this article. The syntactic tree is mainly constructed using the Illinois chunker (Punyakanok and Roth 2001). The kernel used for learning is a preference kernel (Shen and Joshi 2003; Moschitti 2008), which decomposes into Partial Tree Kernel (Moschitti 2006) applications between pairs of enriched tree structures. Word unigram matching is also included in the kernel computation, thus being quite similar to DR-LEX.

Table 8 shows the results obtained on the same WMT12 data set by using only discourse structures, only syntactic structures or both structures together. As we can see, the $\tau$ scores of the syntactic and the discourse variants are not very different (with a general advantage for syntax), but when put together there is a sizeable improvement in correlation for all the language pairs and overall. This is clear evidence that the discourse-based features are providing additional information, which is not included in syntax.

[14] This also explains why translation 12(c) obtains a relatively high score of 0.75. There are many subtree matches coming from the words, and from some very simple and meaningless fragments in the discourse tree (e.g., NUC–Nucleus, NUC–Satellite).

[15] ASIYA syntactic metrics are described on pages 21–24 of the manual (`http://asiya.lsi.upc.edu/`).

**Table 8**
Kendall's (τ) segment level correlation with human judgments on WMT12 obtained by the pairwise preference kernel learning. Results are presented for each language pair and overall.

| **Structure** | CS-EN | DE-EN | ES-EN | FR-EN | Overall |
|---|---|---|---|---|---|
| Syntax | 0.190 | 0.244 | 0.198 | 0.158 | 0.198 |
| Discourse | 0.176 | 0.235 | 0.166 | 0.160 | 0.184 |
| Syntax+Discourse | **0.210** | **0.251** | **0.240** | **0.223** | **0.231** |

## 6. Related Work

In this section we provide a brief overview of related work on discourse in MT (Section 6.1), followed by work on MT evaluation (Section 6.2). In the latter, we cover MT evaluation in general, and in the context of discourse analysis. We also discuss our previous work on using discourse for MT evaluation.

### 6.1 Discourse in Machine Translation

The earliest work on using discourse in machine translation that we are aware of dates back to 2000: Marcu, Carlson, and Watanabe (2000) proposed rewriting discourse trees for MT. However, this research direction was largely ignored by the research community as the idea was well ahead of its time: Note that it came even before the current standard phrase-based SMT model was envisaged (Koehn, Och, and Marcu 2003).

Things have changed since then, and today there is a vibrant research community interested in using discourse for MT, which has started its own biannual Workshop on Discourse in Machine Translation, *DiscoMT* (Webber et al. 2013, 2015; Webber, Popescu-Belis, and Tiedemann 2017). The 2015 edition also started a shared task on cross-lingual pronoun translation (Hardmeier et al. 2015), which had a continuation at WMT 2016 (Guillou et al. 2016), and which is now being featured also at DiscoMT 2017. These shared tasks have the goals of establishing the state of the art and creating common data sets that would help future research in this area.

At this point, several discourse-related research problems have been explored in MT:

- **consistency in translation** (Carpuat 2009; Carpuat and Simard 2012; Ture, Oard, and Resnik 2012; Guillou 2013);
- **lexical and grammatical cohesion and coherence** (Tiedemann 2010a, 2010b; Gong, Zhang, and Zhou 2011; Hardmeier, Nivre, and Tiedemann 2012; Voigt and Jurafsky 2012; Wong and Kit 2012; Ben et al. 2013; Xiong et al. 2013; Louis and Webber 2014; Tu, Zhou, and Zong 2014; Xiong, Zhang, and Wang 2015);
- **word sense disambiguation** (Vickrey et al. 2005; Carpuat and Wu 2007; Chan, Ng, and Chiang 2007);
- **anaphora resolution and pronoun translation** (Hardmeier and Federico 2010; Le Nagard and Koehn 2010; Guillou 2012; Popescu-Belis et al. 2012);

- **handling discourse connectives** (Pitler and Nenkova 2009; Becher 2011; Cartoni et al. 2011; Meyer 2011; Meyer et al. 2012; Meyer and Popescu-Belis 2012; Hajlaoui and Popescu-Belis 2012; Popescu-Belis et al. 2012; Meyer and Poláková 2013; Meyer and Webber 2013; Li, Carpuat, and Nenkova 2014; Steele 2015);
- **full discourse-enabled MT** (Marcu, Carlson, and Watanabe 2000; Tu, Zhou, and Zong 2013).

More details on discourse-related research for MT can be found in the survey (Hardmeier 2012), as well as in the Ph.D. thesis *Discourse in Statistical Machine Translation* (Hardmeier 2014), which received the EAMT Best Thesis Award in 2014.

### 6.2 Discourse in Machine Translation Evaluation

Despite the research interest, so far most attempts to incorporate discourse-related knowledge in MT have been only moderately successful, at best.[16] 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; Meyer 2014). Thus, there is consensus that discourse-informed MT evaluation metrics are needed in order to advance MT research.

The need to consider discourse phenomena in MT evaluation was also emphasized earlier by the Framework for Machine Translation Evaluation in ISLE (FEMTI) (Hovy, King, and Popescu-Belis 2002), which defines quality models (i.e., desired MT system qualities and their metrics) based on the intended context of use.[17] The *suitability* requirement of MT system in the FEMTI comprises discourse aspects including *readability, comprehensibility, coherence, and cohesion*.

In Section 4, we have suggested some simple ways to create such metrics, and we have also shown that they yield better correlation with human judgments. Indeed, we have shown that using linguistic knowledge related to discourse structures can improve existing MT evaluation metrics. Moreover, we have further proposed a state-of-the-art evaluation metric that incorporates discourse information as one of its information sources.

Research in automatic evaluation for MT is very active, and new metrics are constantly being proposed, especially in the context of the MT metric comparisons (Callison-Burch et al. 2007) and metric shared tasks that ran as part of the Workshop on Machine Translation or WMT (Callison-Burch et al. 2008, 2009, 2010, 2011, 2012; Macháček and Bojar 2013, 2014; Stanojević et al. 2015; Bojar et al. 2016), and the NIST Metrics for Machine Translation Challenge, or MetricsMATR.[18] For example, at WMT15, 11 research teams submitted 46 metrics to be compared (Stanojević et al. 2015).

Many metrics at these evaluation campaigns explore ways to incorporate syntactic and semantic knowledge. This reflects the general trend in the field. 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

[16] A notable exception is the work of Tu, Zhou, and Zong (2013), who report up to 2.3 BLEU points of improvement for Chinese-to-English translation using an RST-based MT framework.

[17] `http://www.isi.edu/natural-language/mteval/`.

[18] `http://www.itl.nist.gov/iad/mig/tests/metricsmatr/`.

exploit the similarity over named entities, predicate–argument structures (Giménez and Màrquez 2007; Lo, Tumuluru, and Wu 2012), or semantic frames (Lo and Wu 2011). Finally, there are metrics that combine several lexico-semantic aspects (Giménez and Màrquez 2010b).

As we mentioned earlier, one problem with discourse-related MT research is that it might need specialized evaluation metrics to measure progress. This is especially true for research focusing on relatively rare discourse-specific phenomena, as getting them right or wrong might be virtually "invisible" to standard MT evaluation measures such as BLEU, even when manual evaluation does show improvements (Meyer et al. 2012; Taira, Sudoh, and Nagata 2012; Novák, Nedoluzhko, and Žabokrtský 2013).

Thus, specialized evaluation measures have been proposed, for example, for the translation of discourse connectives (Hajlaoui and Popescu-Belis 2012; Meyer et al. 2012; Hajlaoui 2013) and for pronominal anaphora (Hardmeier and Federico 2010), among others.

In comparison to the syntactic and semantic extensions of MT metrics, there have been very few previous attempts to incorporate discourse information. One example includes the semantics-aware metrics of Giménez and Màrquez (2009) and Giménez et al. (2010), which used the Discourse Representation Theory (Kamp and Reyle 1993) and tree-based discourse representation structures (DRS) produced by a semantic parser. They calculated the similarity between the MT output and the references based on DRS subtree matching as defined in Liu and Gildea (2005), also using DRS lexical overlap and DRS morpho-syntactic overlap. However, they could not improve correlation with human judgments as evaluated on the MetricsMATR data set, which consists of 249 manually assessed segments. Compared with that previous work, here (i) we used a different discourse representation (RST), (ii) we compared discourse parses using *all-subtree* kernels (Collins and Duffy 2001), (iii) we evaluated on much larger data sets, for several language pairs and for multiple metrics, and (iv) we did demonstrate better correlation with human judgments.

Recently, other discourse-related extensions of MT metrics (such as BLEU, TER, and METEOR) were proposed (Wong et al. 2011; Wong and Kit 2012), which use document-level **lexical cohesion** (Halliday and Hasan 1976). In that work, 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 data set (900 segments) when linearly interpolating these metrics with their lexical cohesion score. However, they ignored a key property of discourse, namely, the coherence structure, which we have effectively exploited in both tuning and no-tuning scenarios. Furthermore, we have shown that the similarity in discourse trees can yield improvements in a larger number of existing MT evaluation metrics. Finally, unlike their work, which measured lexical cohesion at the document level, here we are concerned with coherence (rhetorical) structure, primarily at the sentence level.

Finally, we should note our own previous work, on which this article is based. In Guzmán et al. (2014b), we showed that using discourse can improve a number of pre-existing evaluation metrics, and in Joty et al. (2014) we presented our DiscoTK family of discourse-based metrics. In particular, the DISCOTK$_{party}$ metric (discussed in Section 4.3) combined several variants of a discourse tree representation with other metrics from the ASIYA MT evaluation toolkit, and yielded the best-performing metric in the WMT14 Metrics shared task. Compared with those previous publications of ours, here we provide additional detail and extensive analysis, trying to explain why discourse information is helpful for MT evaluation.

In another related publication (Guzmán et al. 2014a), we proposed a pairwise learning-to-rank approach to MT evaluation that learns to differentiate better from worse translations compared with a given reference. There, we integrated several layers of linguistic information, combing POS, shallow syntax, and discourse parse, which we encapsulated in a common tree-based structure.

We used preference re-ranking kernels to learn the features automatically. The evaluation results show that learning in the proposed framework yields better correlation with human judgments than computing the direct similarity over the same type of structures. Also, we showed that the structural kernel learning can be a general framework for MT evaluation, in which syntactic and semantic information can be naturally incorporated.

Unfortunately, learning features with preference kernels is computationally very expensive, both at training and at testing time. Thus, in a subsequent work (Guzmán et al. 2015), we used a pairwise neural network instead, where lexical, syntactic, and semantic information from the reference and the two hypotheses is compacted into small distributed vector representations and fed into a multilayer neural network that models the interaction between each of the hypotheses and the reference, as well as between the two hypotheses. This framework yielded correlation with human judgments that rivals the state of the art. In future work, we plan to incorporate discourse information in this neural framework, which we could not do initially because of the lack of discourse embeddings. However, with the availability of a neural discourse parser like the one proposed by Li, Li, and Hovy (2014), this goal is now easily achievable.

## 7. Conclusions

We addressed the research question of whether sentence-level discourse structure can help the automatic evaluation of machine translation. To do so, we defined several variants of a simple discourse-aware similarity metric, which use the all-subtree kernel to compute similarity between RST trees. We then used this similarity metric to automatically assess MT quality in the evaluation benchmarks from the WMT metrics shared task. We proposed to take the similarity between the discourse trees for the hypothesis and for the reference translation as an absolute measure of translation quality. The results presented here can be analyzed from several perspectives:

*Applicability.* The first conclusion after a series of experimental evaluations is that the sentence-level discourse structure can be successfully leveraged to evaluate translation quality. Although discourse-based metrics perform reasonably well on their own, especially at the system level, one interesting fact is that discourse information is complementary to many existing metrics for MT evaluation (e.g., BLEU, TER, METEOR) that encompass different levels of linguistic information. At a system level, this leads to systematic improvements in correlation with human judgments in the majority of the cases where the discourse-based metrics were mixed with other single metrics in a uniformly weighted linear combination. When we further tuned the combination weights via supervised learning from human-assessed pairwise examples, we obtained even better results and observed average relative gains between 22% and 35% in segment-level correlations.

*Robustness.* Other interesting properties we observed in our experiments have to do with the robustness of the supervised learning combination approach. The results were very stable when training and testing across several WMT data sets from different years.

Additionally, the tuned metrics were quite insensitive to the source language in the translation, to the point that it was preferable to train with all training examples together rather than training source-language specific models.

*External validation.* Exploiting this combination approach to its best, we produced a strong combined MT evaluation metric (DISCOTK$_{party}$) composed by 23 individual metrics, including five variants of our discourse-based metric, which performed best at the WMT14 translation evaluation task, both at the system level and at the segment level. When building the state-of-the-art DISCOTK$_{party}$ metric, we observed that discourse-based features are favorably weighted (e.g., DR-LEX$_1$ was ranked fourth out of the 23 metrics), having coefficients that are on par with other features such as BLEU. This tells us that the contribution of discourse-based information is significant even in the presence of such a rich diversity of sources of information. In this direction, we also presented evidence showing that the contribution from the sentence-level discourse information is beyond what the syntactic information provides to the evaluation metrics; in fact, the two linguistic dimensions collaborate well, producing cumulative gains in performance.

*Understanding the Contribution of Discourse.* In this article, we also presented a more qualitative analysis in order to better understand the contribution of the discourse trees in the new proposed discourse metrics. First, we conducted an ablation study, and we confirmed that all layers of information present in the discourse trees (i.e., hierarchical structure, discourse relations, and nuclearity labels) play a role and make a positive incremental contribution to the final performance. Interestingly, the most relevant piece of information is the nuclearity labels, rather than the relations.

Second, we analyzed the ability of discourse trees to discriminate between good and bad translations in practice. For that, we computed the similarity between the discourse trees of the reference translations and the discourse trees of a set of *good* translations, and compared it with the similarity between the discourse trees of the reference translations and a set of *bad* translations. *Good* and *bad* translations were selected based on existing human evaluations. This similarity was computed at different levels, including relation labels, nuclearity labels, elementary discourse units, tree depth, and so forth. We observed a systematically higher similarity to the discourse trees of *good* translations, in all the specific elements tested and across all language pairs. These results confirm the ability of discourse trees to characterize *good* translations as the ones more similar to the reference.

*Limitations and Future Work.* An important limitation of our study is that it is restricted to sentence-level discourse parsing. Although it is true that complex sentences with non-trivial discourse structure abound in our corpora, it is reasonable to think that there is more potential in the application of discourse parsing at the paragraph or at the document level. The main challenge in this direction is that there are no corpora available with manual annotations of the translation quality at the document level.

Second, we have applied discourse only for MT evaluation, but we would like to follow a similar path to verify whether discourse can also help machine translation itself. Our first approach will be to use discourse information to re-rank a set of candidate translations. The main challenge here is that one has to establish the links between the discourse structure of the source and that of the translated sentences, trying to promote translations that preserve discourse structure.

Finally, at the level of learning, we are working on how to jump from tuning the overall weights of a linear combination of metrics to perform learning on fine-grained features, for example, consisting of the substructures that the discourse parse tree, and other linguistic structures (syntax, semantics, etc.) contain. This way, we would be learning the features that help in identifying better translations compared to worse translations (Guzmán et al. 2014a; Guzmán et al. 2015). Our vision is to have a model that can learn combined evaluation metrics taking into account different levels of linguistic information, fine-grained features, and pre-existing measures, and which could be applied, with minor variations, to the related problems of MT evaluation, quality estimation, and reranking.

## Acknowledgments

The authors would like to thank the reviewers of the previous versions of this article for their constructive and thorough comments and criticism. Thanks to their comments we have improved the article significantly.

## References

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

Bazrafshan, Marzieh and Daniel Gildea. 2014. Comparing representations of semantic roles for string-to-tree decoding. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1786–1791, Doha.

Becher, Viktor. 2011. When and why do translators add connectives? A corpus-based study. *Target. International Journal of Translation Studies*, 23(1):26–47.

Ben, Guosheng, Deyi Xiong, Zhiyang Teng, Yajuan Lü, and Qun Liu. 2013. Bilingual lexical cohesion trigger model for document-level machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 382–386, Sofia.

Bojar, Ondřej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, MD.

Bojar, Ondřej, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation*, pages 199–231, Berlin.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, OH.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala.

Callison-Burch, Chris, 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.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens.

Callison-Burch, Chris, 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.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento.

Carlson, Lynn and Daniel Marcu. 2001, Discourse tagging reference manual. Technical Report ISI-TR-545, University of Southern California Information Sciences Institute.

Carpuat, Marine. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27, Boulder, CO.

Carpuat, Marine and Michel Simard. 2012. The trouble with SMT consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 442–449, Montreal.

Carpuat, Marine and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 61–72, Prague.

Cartoni, Bruno, 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, OR.

Chan, Yee Seng, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague.

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

Chiang, David, 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*, Association for Computational Linguistics, Honolulu, HI.

Collins, Michael and Nigel Duffy. 2001. Convolution kernels for natural language. In *Neural Information Processing Systems*, pages 625–632, Vancouver.

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

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

Danlos, Laurence. 2009. D-STAG: A discourse analysis formalism based on synchronous tags. *TAL*, 50(1):111–143.

Denkowski, Michael 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.

Denkowski, Michael 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*, San Diego, CA.

Doddington, George. 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*, pages 138–145, Morgan Kaufmann Publishers Inc., San Diego, CA.

Galley, Michel, 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*, pages 273–280, San Diego, CA.

Giménez, Jesús 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.

Giménez, Jesús 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.

Giménez, Jesús and Lluís Màrquez. 2010a. Asiya: An open toolkit for automatic machine translation (meta-)evaluation. *The Prague Bulletin of Mathematical Linguistics*, 94:77–86.

Giménez, Jesús and Lluís Màrquez. 2010b. Linguistic measures for automatic machine translation evaluation. *Machine Translation*, 24(1):77–86.

Giménez, Jesús, Lluís Màrquez, Elisabet Comelles, 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.

Gong, Zhengxian, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 909–919, Edinburgh.

Gonzàlez, Meritxell, Jesús Giménez, and Lluís Màrquez. 2012. A graphical interface for MT evaluation and error analysis. In *Proceedings of the ACL 2012 System Demonstrations*, pages 139–144, Jeju Island.

Guillou, Liane. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon.

Guillou, Liane. 2013. Analysing lexical consistency in translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 10–18, Sofia.

Guillou, Liane, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*, pages 525–542, Berlin.

Guzmán, Francisco, Shafiq Joty, Lluís Màrquez, Alessandro Moschitti, Preslav Nakov, and Massimo Nicosia. 2014a. Learning to differentiate better from worse translations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 214–220, Doha.

Guzmán, Francisco, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014b. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 687–698, Baltimore, MD.

Guzmán, Francisco, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise neural machine translation evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 805–814, Beijing.

Hajlaoui, Najeh. 2013. Are ACT's scores increasing with better translation quality? In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 408–413, Sofia.

Hajlaoui, Najeh and Andrei Popescu-Belis. 2012. Translating English discourse connectives into Arabic: a corpus-based analysis and an evaluation metric. In *Fourth Workshop on Computational Approaches to Arabic Script-based Languages at Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas*.

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

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

Hardmeier, Christian. 2014. *Discourse in Statistical Machine Translation*. Ph.D. thesis, Uppsala University, Uppsala, Sweden.

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

Hardmeier, Christian, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the EMNLP 2015 Workshop on Discourse in Machine Translation*, Lisbon.

Hardmeier, Christian, 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*, pages 1179–1190, Jeju Island.

Hobbs, Jerry. 1979. Coherence and Coreference. *Cognitive Science*, 3:67–90.

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

Hovy, Eduard, Margaret King, and Andrei Popescu-Belis. 2002. Principles of context-based machine translation evaluation. *Machine Translation*, 17(3–4):43–75.

Joty, Shafiq, 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*, pages 904–915, Jeju Island.

Joty, Shafiq, Giuseppe Carenini, and Raymond T. Ng. 2015. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. *Computational Linguistics*, 41(3):385–435.

Joty, Shafiq, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2014. DiscoTK: Using discourse structure for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 402–408, Baltimore, MD.

Kamp, Hans and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Model Theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. 42. Kluwer Academic Publishers.

Kendall, Maurice. 1938. A new measure of rank correlation. *Biometrika*, 1–2(30):81–89.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Edmonton, Canada.

Kullback, Solomon and Richard A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Lavie, Alon 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*, pages 228–231, Prague.

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

Le Nagard, Ronan and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala.

Li, Jiwei, Rumeng Li, and Eduard Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069, Doha.

Li, Junyi Jessy, Marine Carpuat, and Ani Nenkova. 2014. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 283–288, Baltimore, MD.

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization Branches Out*, pages 74–81, Barcelona.

Liu, Ding 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, MI.

Lo, Chi kiu, 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.

Lo, Chi kiu and Dekai Wu. 2011. Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 220–229, Portland, OR.

Louis, Annie and Bonnie Webber. 2014. Structured and unstructured cache models for SMT domain adaptation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 155–163, Gothenburg.

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

Macháček, Matouš and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, MD.

Mann, William and Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Marcu, Daniel, Lynn Carlson, and Maki Watanabe. 2000. The automatic translation of discourse structures. In *Proceedings of the 1st North American Chapter of the Association* *for Computational Linguistics Conference*, pages 9–17, Seattle, WA.

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

Meyer, Thomas. 2014. *Discourse-level Features for Statistical Machine Translation*. Ph.D. thesis, École polytechnique Fédérale de Lausanne (EPFL), Lausanne.

Meyer, Thomas and Lucie Poláková. 2013. Machine translation with many manually labeled discourse connectives. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 43–50, Sofia.

Meyer, Thomas and Andrei Popescu-Belis. 2012. Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 129–138, Avignon.

Meyer, Thomas, 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)*, San Diego, CA.

Meyer, Thomas and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, Sofia.

Moschitti, Alessandro. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *17th European Conference on Machine Learning*, pages 318–329, Berlin.

Moschitti, Alessandro. 2008. Kernel methods, syntax and semantics for relational text categorization. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 253–262, Napa Valley, CA.

Nadejde, Maria, Philip Williams, and Philipp Koehn. 2013. Edinburgh's syntax-based machine translation systems. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 170–176, Sofia.

Novák, Michal, Anna Nedoluzhko, and Zdeněk Žabokrtský. 2013. Translation of "it" in a deep syntax framework. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 51–59, Sofia.

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

Papineni, Kishore, 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*, Philadelphia, PA.

Pearson, Karl. 1895. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, (58):240–242.

Pitler, Emily and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing Conference*, pages 13–16, Suntec.

Popescu-Belis, Andrei, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni, and Sandrine Zufferey. 2012. Discourse-level annotation over Europarl for machine translation: Connectives and pronouns. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2716–2720, Istanbul.

Popovic, Maja 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.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 921–950, Marrakech.

Punyakanok, Vasin and Dan Roth. 2001. The use of classifiers in sequential inference. In *Advances in Neural Information Processing Systems 14*, pages 995–1001, Vancouver.

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

Shen, Libin and Aravind K. Joshi. 2003. An SVM-based voting algorithm with application to parse reranking. In

*Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 9–16, Edmonton.

Smola, Alex and S.v.n. Vishwanathan. 2003. Fast kernels for string and tree matching. In *Advances in Neural Information Processing Systems 15*, pages 585–592, Vancouver.

Snover, Matthew, 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)*, pages 223–231, Cambridge, MA.

Spearman, Charles. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101.

Stanojević, Miloš, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon.

Steele, David. 2015. Improving the translation of discourse markers for Chinese into English. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 110–117, Denver, CO.

Sutton, Charles, 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.

Taboada, Maite and William C. Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

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

Taira, Hirotoshi, Katsuhito Sudoh, and Masaaki Nagata. 2012. Zero pronoun resolution can improve the quality of j-e translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 111–118, Jeju Island.

Tiedemann, Jörg. 2010a. Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15, Uppsala.

Tiedemann, Jörg. 2010b. To cache or not to cache? Experiments with adaptive models in statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 189–194, Uppsala.

Tu, Mei, Yu Zhou, and Chengqing Zong. 2013. A novel translation framework based on rhetorical structure theory. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 370–374, Sofia.

Tu, Mei, Yu Zhou, and Chengqing Zong. 2014. Enhancing grammatical cohesion: Generating transitional expressions for SMT. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 850–860, Baltimore, MD.

Ture, Ferhan, Douglas W. Oard, and Philip Resnik. 2012. Encouraging consistent translation choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 417–426, Montréal.

Vickrey, David, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 771–778, Vancouver.

Voigt, Rob and Dan Jurafsky. 2012. Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montréal.

Watanabe, Taro, 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*, pages 764–773, Prague.

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

Webber, Bonnie, Marine Carpuat, Andrei Popescu-Belis, and Christian Hardmeier, editors. 2015. *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Lisbon.

Webber, Bonnie, Andrei Popescu-Belis, Katja Markert, and Jörg Tiedemann, editors. 2013. *Proceedings of the Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Sofia.

Webber, Bonnie, Andrei Popescu-Belis, and Jörg Tiedemann, editors. 2017. *Proceedings of the Third Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Copenhagen.

Wong, Billy T. M. 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*, pages 1060–1068, Jeju Island.

Wong, Billy T. M., Cecilia F. K. Pun, Chunyu Kit, and Jonathan J. Webster. 2011. Lexical cohesion for evaluation of machine translation at document level. In *Proceedings of the 7th International Conference on Natural Language Processing and Knowledge Engineering*, pages 238–242, Tokushima.

Wu, Dekai 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*, pages 13–16, Boulder, CO.

Xiong, Deyi, Yang Ding, Min Zhang, and Chew Lim Tan. 2013. Lexical chain based cohesion models for document-level statistical machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1563–1573, Seattle, WA.

Xiong, Deyi, Min Zhang, and Xing Wang. 2015. Topic-based coherence modeling for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):483–493.

Yeh, Alexander. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, pages 947–953, Saarbrücken.