Discourse and Document-level Information for Evaluating Language Output Tasks

Carolina Scarton
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
c.scarton@sheffield.ac.uk

Abstract

Evaluating the quality of language output tasks such as Machine Translation (MT) and Automatic Summarisation (AS) is a challenging topic in Natural Language Processing (NLP). Recently, techniques focusing only on the use of outputs of the systems and source information have been investigated. In MT, this is referred to as Quality Estimation (QE), an approach that uses machine learning techniques to predict the quality of unseen data, generalising from a few labelled data points. Traditional QE research addresses sentence-level QE evaluation and prediction, disregarding document-level information. Document-level QE requires a different set up from sentence-level, which makes the study of appropriate quality scores, features and models necessary. Our aim is to explore document-level QE of MT, focusing on discourse information. However, the findings of this research can improve other NLP tasks, such as AS.

1 Introduction

Evaluation metrics for Machine Translation (MT) and Automatic Summarisation (AS) tasks should be able to measure quality with respect to different aspects (e.g. fluency and adequacy) and they should be fast and scalable. Human evaluation seems to be the most reliable (although it might introduce biases of reviewers). However, it is expensive and cumbersome for large datasets; it is also not practical for certain scenarios, such as *gisting* in MT and summarisation of webpages.

Automatic evaluation metrics (such as BLEU (Papineni et al., 2002) and ROUGE (Lin and Och, 2004)), based on human references, are widely used to evaluate MT and AS outputs. One limitation of these metrics is that if the MT or AS system outputs a translation or summary considerably different from the references, it does not really mean that it is a bad output. Another problem is that these metrics cannot be used in scenarios where the output of the system is to be used directly by end-users, for example a user reading the output of Google Translate[1] for a given news text cannot count on a reference for that translated text.

Quality Estimation (QE) approaches aim to predict the quality of MT systems without using references. Instead, features (that may be or may not be related to the MT system that produced this translations) are applied to source and target documents (Blatz et al., 2004; Bojar et al., 2013). The only requirement is data points with scores (e.g.: Human-targeted Translation Error Rate (HTER) (Snover et al., 2006) or even BLEU-style metrics). These data points can be used to train supervised machine learning models (regressors or classifiers) to predict the scores of unseen data. The advantage of these approaches is that we do not need to have all the words, sentences or documents of a task evaluated manually, we just need enough data points to train the machine learning model.

QE systems predict scores that reflect how good a translation is for a given scenario. For example, a widely predicted score in QE is HTER, which measures the effort needed to post-edit a sentence. A

[1] https://translate.google.com/

Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), pages 118–125,
Denver, Colorado, June 1, 2015.

user of a QE system predicting HTER could decide whether to post-edit or translate sentences from scratch based on the score predicted for each sentence.

The vast majority of work done on QE is at sentence level. Document-level predictions, on the other hand, are interesting in scenarios where one wants to evaluate the overall score of an MT system or where the end-user is interested in the quality of the document as whole. In addition, document-level features can also correlate well with quality scores, mainly because state-of-the-art MT systems translate documents at sentence level, disregarding discourse information. Therefore, it is expected that the outputs of these systems may contain discourse problems.

In this work we focus on document-level QE. Regarding features, discourse phenomena are being considered since they are linguistic phenomena that often manifest document-wide. These phenomena are related to how sentences are connected, how genre and domain of a document are identified, anaphoric pronouns, etc.

Regarding document-level prediction, we focus on finding the ideal quality label for the task. Traditional evaluation metrics tend to yield similar scores for different documents. This leads to low variation between the document quality scores with all these scores being close to the mean score. Therefore, a quality label that captures document quality in a more sensitive way is needed.

Research on the use of linguistic features for QE and the use of discourse for improving MT and MT evaluation are presented in Section 2. Section 3 presents the work done so far and the directions that we intend to follow. Conclusions and future work are presented in Section 4

2 Document-level information for QE and MT

Traditional systems translate documents at sentence level, disregarding document-wide information. This means that sentences are translated without considering the relations in the whole document. Therefore, information such as discourse structures can be lost in this process.

QE is also traditionally done at sentence level mainly because the majority of MT systems translate texts at this level. Another reason is that sentence-level approaches have more applications than other granularity levels, because they can explore the peculiarities of each sentence, being very useful for the post-edition task. On the other hand, sentence-level approaches do not consider the document as a whole and information regarding discourse is disregarded. Moreover, for scenarios in which post-edition is not possible, for example, *gisting*, quality predictions for the entire documents are more useful.

In this section we present related work on QE and the first research towards document-level QE. Research on the use of discourse phenomena for MT improvement and MT evaluation are also presented.

2.1 Quality Estimation of Machine Translation

Previous work on QE has used supervised machine learning (ML) approaches (mainly regression algorithms). Besides the specific ML method adopted, the choice of features is also a design decision that plays a crucial role.

Sentences (or documents) from source and target and also information from the MT system are used for designing features. The features extracted are used as input to train a QE model. In this training phase supervised ML techniques, such as regression, can be applied. A training set with quality labels is provided for an ML model. These quality labels are the scores that the QE model will learn to predict. Therefore, the QE model will be able to predict a quality score for a new, unseen data points. The quality labels can be *likert* scores, HTER, BLEU, just to cite some widely used examples. Also the ML algorithm can vary (SVM and Gaussian Process are the state-of-the-art algorithms for QE).

Some work in the area include linguistic information as features for QE (Avramidis et al., 2011; Pighin and Màrquez, 2011; Hardmeier, 2011; Felice and Specia, 2012; Almaghout and Specia, 2013) at sentence level. Only Scarton and Specia (2014) (predicting quality at document level) and Rubino et al. (2013) (sentence level) focus on the use of discourse information for QE.

It is important to notice that frameworks like QuEst[2] (Specia et al., 2013) are available for QE at

[2]http://www.quest.dcs.shef.ac.uk

sentence level. QuEst has modules to extract several features for QE from source and target documents and to experiment with ML techniques for predicting QE. Features are divided in two types: glass-box (dependent on the MT system) and black-box (independent on the MT system).

At document level, Soricut and Echihabi (2010) explore document-level QE prediction to rank documents translated by a given MT system, predicting BLEU scores. Features include text-based, language model-based, pseudo-reference-based, example-based and training-data-based. Pseudo-reference features are BLEU scores based on pseudo-references from an off-the-shelf MT system, for both the target and the source languages.

Scarton and Specia (2014) explore lexical cohesion and LSA (Latent Semantic Analysis) (Landauer et al., 1998) cohesion for document-level QE. The lexical cohesion features are repetitions (Wong and Kit, 2012) and the LSA cohesion is achieved following the work of Graesser et al. (2004). Pseudo-reference features are also applied in this work, according to the work of Soricut and Echihabi (2010). BLEU and TER (Snover et al., 2006) are used as quality labels. The best results were achieved with pseudo-reference features. However, LSA cohesion features alone also showed improvements over the baseline.

2.2 Discourse phenomena in MT

In the MT area, there have been attempts to use discourse information that can be used as inspiration source for QE features. The need of document-level information for improving MT is a widely accepted fact. However, it is hard to integrate discourse information into traditional state-of-the-art sentence-level MT systems. It is also challenging to build a document-level or discourse-based MT system from scratch. Therefore, the initiatives focus on the integration of discourse as features into the decoding phase or previously annotate discourse phenomena in the parallel corpora.

Lexical Cohesion is related to word usage: word repetitions, synonyms repetitions and collocations. Besides initiatives to improve MT system and outputs with lexical cohesion (Ture et al., 2012; Xiao et al., 2011; Ben et al., 2013), Wong and Kit (2012) apply lexical cohesion metrics for evaluation of MT systems at document level.

Coreference is related to coherence clues, such as pronominal anaphora and connectives. Machine translation can break coreference chains since it is done at sentence level. Initiatives for improvement of coreference in MT include anaphora resolution (Giménez et al., 2010; LeNagard and Kohen, 2010; Hardmeier and Federico, 2010; Hardmeier, 2014) and connectives (Popescu-Belis et al., 2012; Meyer and Popescu-Belis, 2012; Meyer et al., 2012; Li et al., 2014).

RST (Rhetorical Structure Theory) (Mann and Thompson, 1987) is a linguistic theory that correlates macro and micro units of discourse in a coherent way. The correlation is made among EDUs (Elementary Discourse Units). EDUs are defined at sentence, phrase or paragraph-level. These correlations are represented in the form of a tree. Marcu et al. (2000) explore RST focusing on identifying the feasibility of building a discourse-based MT system. Guzmán et al. (2014) use RST trees comparison for MT evaluation.

Topic models capture word usage, although they are more robust than lexical cohesion structures because they can correlate words that are not repetitions or do not present any semantic relation. These methods can measure if a document follows a topic, is related to a genre or belongs to a specific domain. Work on improving MT that uses topic models include Zhengxian et al. (2010) and Eidelman et al. (2012).

3 Planned Work

In this paper, we describe the three main research questions that we aim to answer in this PhD work:

1. How to address document-level QE?
2. Are discourse models appropriate to be used for QE at document level? Are these models applicable for different languages?
3. How can we use the discourse information for the evaluation of Automatic Summarisation and Readability Assessment?

In this section, we summarise how we are addressing these research questions.

3.1 Document-level Quality Estimation

As mentioned previously, one aim of this PhD is to identify a suitable quality label for document-level QE. Our hypothesis is that document quality is more complex than a simple aggregation of sentence quality. In order to exemplify this assumption, consider document A and document B. Documents A and B have the same number of sentences (10 sentences) and score the same value when we access quality as an average of HTER at sentence level, 0.5. However, 5 sentences of document A score 1 and the other five sentences score 0. On the other hand, document B shows a more smooth distribution of scores among sentences (the majority of the sentences score a value close to 0.5). Are document A and B comparable just because the averaged HTERs are the same? Our assumption is that a real score at document level or a more clever combination of sentence-level scores are the more suitable ways to evaluate documents.

Another drawback of averaging sentence-level scores is that sentences have different importance inside a document, they contain different information across a document. Therefore, documents that have important sentences badly translated should be penalised more heavily. The way we propose to address this problem is by using summarisation or information retrieval techniques in order to identify the most important sentences (or even paragraphs) and assign different weights according to the relevance of the sentence.

Moreover, we studied several traditional evaluation metrics as quality labels for QE at document level and found out that, on average, all the documents seem to be similar. Part of this study is showed in Table 1 for 9 documents of WMT2013 QE shared task corpus (English-Spanish translations) and for 119 documents of LIG corpus (Potet et al., 2012) (French-English translations, with post-editions).[3] The quality metrics considered were BLEU, TER, METEOR (Banerjee and Lavie, 2005) and an average of HTER scores at sentence level.

All traditional MT evaluation metrics showed low standard deviation (STDEV) in both corpora. Also the HTER at sentence level averaged to obtain a document-score showed low variation. This means that all documents in the corpora seem similar in terms of quality. Our hypothesis is that this evaluation is wrong and other factors should be considered in order to achieve a suitable quality label for document-level prediction.

Besides quality scores, another issue in document-level QE is the features to be used. Thus far, the majority of features for QE are at word or sentence level. Since a document can be viewed as a combination of words and sentences one way to explore document-level features is to combine word- or sentence-level features (by averaging them, for example). Another way is to explore linguistic phenomena document-wide. This is discussed on the next subsection.

New features and prediction at document level can be included in existing frameworks, such as QuEst. This is the first step to integrate document-level and sentence-level prediction and features.

3.2 Modelling discourse for Quality Estimation

Discourse phenomena happen document-wide and, therefore, these can be considered a strong candidate for the extraction of document-level features. A document is not only a bag of words and sentences, although the words and sentences are in fact organised in a logical way by using linguistic clues. Discourse was already studied in the MT field, aiming to improve MT systems and/or MT outputs and also to automatically evaluate MT against human references. However, for QE, we should be able to deal with evaluation for several language pairs, considering features for source and target. Another issue is that QE features should correlate with the quality score used. Therefore, the use of discourse for QE purposes deserves further investigation.

We intend to model discourse for QE by applying linguistic and statistical knowledge. Two cases are being explored:

3.2.1 Linguistic-based models

Certain discourse theories could be used to model discourse for QE purposes, such as such as the Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) and Entity-Grid models (Barzilay and Lapata, 2008; Elsner, 2011). We refer to these two theories mainly because they can be readily applied, for English language, given the existence of

[3] Both corpora were translated by only one SMT system.

	WMT		LIG	
	Average	STDEV	Average	STDEV
BLEU (↑)	0.26	0.046	0.27	0.052
TER (↓)	0.52	0.049	0.53	0.069
METEOR-ex (↑)	0.46	0.050	0.29	0.031
METEOR-st (↑)	0.43	0.050	0.30	0.030
Averaged HTER (↓)	0.25	0.071	0.21	0.032

Table 1: Average values of evaluation metrics in the WMT and LIG corpora

parsers (RST parser (Joty et al., 2013) and Entity Grid parser).[4] Although these resources are only available for English, it is important in this stage to study the impact of this information for document-level QE, considering English as source or target language. In this scenario, we intend to explore source and target features isolated (source features will be applied only when English is source language and target features only when English is target).

Moreover, other linguistic information could be used to model discourse for QE. Anaphoric information, co-reference resolution and discourse connectives classification could be used. (Scarton and Specia, 2014) explore lexical cohesion features for QE. These features are based on repetitions of words or lemmas. Looking at more complex structures, such as synonym in order to count repetitions beyond word matching can lead to improvements in the results.

We have also studied linguistic phenomena and their correlations with HTER values at document level on the LIG corpus. Results are shown in Figure 1. This figure shows four scenarios with different numbers of documents. The first scenario has ten documents: the five best documents and the five worst (in terms of averaged HTER). The second scenario considers the ten best and ten worst, the third the 20 best and 20 worst and the fourth the 40 best and 40 worst. The last scenario considers all the data. The bars are Pearson's r correlation values between a given feature and the real HTER value. Features were: number of connectives, number of pronouns, number of RST nucleus relations, number of RST satellite relations, number of elementary discourse units (EDUs) breaks, lexical cohesion (LC) features and LSA features from the work of (Scarton and Specia, 2014). The most successful features of QuEst framework were also considered: QuEst1 - number of tokens, QuEst2 - language model probability, QuEst3 - number of occurrences of the target word within the target hypothesis (averaged for all words in the hypothesis - type/token ratio) and QuEst4 - number of punctuation marks. Features were only applied for target (English) due to resources availability.

Mainly for scenarios with 10 and 20 documents, features considering discourse phenomena counts performed better than QuEst features. In the other scenarios LC and LSA features were the best. It is worth mentioning that this study is on-going and much more can be extracted from discourse information such as RST, than only simple counts.

3.2.2 Latent variable models

Latent variable models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Latent Semantic Analysis (LSA) (Landauer et al., 1998) have been widely used to extract topic models from corpora. The idea behind these methods is that a matrix of words versus sentences is built and mathematical transformations are applied, in order to achieve correlations among the word vectors. However, as Graesser et al. (2004) suggests, they can also be used to find lexical cohesion information within documents. In fact, topic modelling approaches have already been used to improve MT and also for QE at sentence level. Their advantage is that they are fast, language independent and do not require robust resources (such as discourse parsers). Previous work has used LSA and LDA for QE purposes (Scarton and Specia, 2014; Rubino et al., 2013).

We could also use latent variable models to find how close a machine translated document is from original documents in the same language, genre and domain.

[4] https://bitbucket.org/melsner/browncoherence

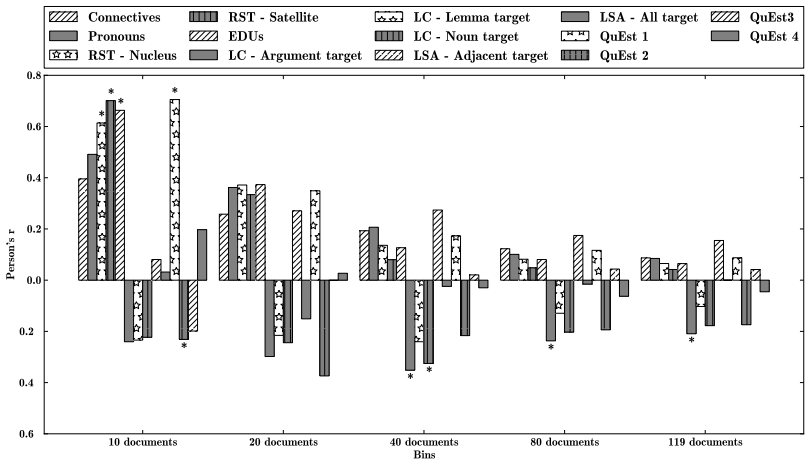

Figure 1: Impact of discourse features on document-level QE - '*' means *p-value* < 0.05

3.3 Using discourse models for other NLP tasks

One of our aims is to evaluate whether the discourse models built for QE can be used for the evaluation of other tasks in NLP. AS evaluation could benefit from QE: to an extent, AS outputs could be viewed as "translations" from "source language" into "source summarised language". Up to now, only (Louis and Nenkova, 2013) proposed an approach for evaluating summaries without references (by using pseudo-references). Moreover, discourse evaluation of AS outputs is expected to show more correlation with quality scores than MT because of the nature of the tasks. While MT outputs are dependent on the source language (and, as shown by Carpuat and Simard (2012), they tend to preserve discourse constituents of the source), AS outputs are built by choosing sentences from one or more documents trying to keep as much relevant information as possible. The combination of text from multiple documents can lead to loss of coherence of automatic summaries more than MT does to translated texts.

Another task in NLP that could benefit from advances in QE is **Readability Assessment (RA)**. This task consists in evaluating the complexity of documents for a given audience (therefore, the task is an evaluation per se). Several studies have already explored discourse information for RA (Graesser et al., 2004; Pitler and Nenkova, 2008; Todirascu et al., 2013). QE techniques can benefit RA in scenarios where we need to compare texts produced by or for native speakers or second language learners (SLL) or texts produced by or for mentally impaired patient compared to healthy subjects (in these scenarios, the documents produced by or for the "experts" could be considered as source documents and documents produced by or for "inexpert or mentally impaired" as target documents).

4 Conclusion

In this paper we presented a proposal to address to document-level quality estimation. This includes the study of quality labels for document-level prediction and also document-level features. We intend to focus on discourse features, because of the nature of discourse phenomena.

We showed that traditional MT evaluation metrics are not suitable for QE at document level because they cannot measure quality of documents according to relevance of sentences.

Discourse features were also evaluated for document-level QE showing higher correlation with HTER scores than the most successful features from QuEst framework. This is sign that discourse information can help in document-level prediction.

Finally, we discussed ways to use the discourse models developed for QE to improve evaluation of other NLP task: AS and RA.

Acknowledgments

This work was supported by the EXPERT (EU Marie Curie ITN No. 317471) project.

References

Hala Almaghout and Lucia Specia. 2013. A CCG-based Quality Estimation Metric for Statistical Machine Translation. In *The XIV Machine Translation Summit*, pages 223–230, Nice, France.

Eleftherios Avramidis, Maja Popovic, David Vilar Torres, and Aljoscha Burchardt. 2011. Evaluate with Confidence Estimation: Machine ranking of translation outputs using grammatical features. In *The Sixth Workshop on Statistical Machine Translation*, pages 65–70, Edinburgh, UK.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *The ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Harbor, MI.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Gousheng Ben, Deyi Xiong, Zhiyang Teng, Yajuan Lu, and Qun Liu. 2013. Bilingual Lexical Cohesion Trigger Model for Document-Level Machine Translation. In *The 51st Annual Meeting of the Association for Computational Linguistics*, pages 382–386, Sofia, Bulgaria.

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

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning research*, 3:993–1022.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *The Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

Marine Carpuat and Michel Simard. 2012. The Trouble with SMT Consistency. In *The Seventh Workshop on Statistical Machine Translation*, pages 442–449, Montréal, Quebec, Canada.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic Models of Dynamic Translation Model Adaptation. In *The 50th Annual Meeting of the Association for Computational Linguistics*, pages 115–119, Jeju Island, Korea.

Micha Elsner. 2011. *Generalizing Local Coherence Modeling*. Ph.D. thesis, Department of Computer Science, Brown University, Providence, Rhode Island.

Mariano Felice and Lucia Specia. 2012. Linguistic Features for Quality Estimation. In *The Seventh Workshop on Statistical Machine Translation*, pages 96–103, Montréal, Quebec, Canada.

Jesús Giménez, Lluís Màrquez, Elisabet Comelles, Irene Catellón, and Victoria Arranz. 2010. Document-level Automatic MT Evaluation based on Discourse Representations. In *The Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 333–338, Uppsala, Sweden.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 36:193–202.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using Discourse Structure Improves Machine Translation Evaluation. In *The 52nd Annual Meeting of the Association for Computational Linguistics*, pages 687–698, Baltimore, MD.

Christian Hardmeier and Marcello Federico. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In *The 7th International Workshop on Spoken Language Translation*, pages 283–289.

Christian Hardmeier. 2011. Improving machine translation quality prediction with syntatic tree kernels. In *Proceedings of the 15th conference of the European Assoaciation for Machine Translation (EAMT 2011)*, pages 233–240, Leuven, Belgium.

Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*. Ph.D. thesis, Department of Linguistics and Philology, Uppsala University, Sweden.

Shafiq Joty, Giuseppe Carenini, Raymond T. Ng, and Yashar Mehdad. 2013. Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis. In *The 51st Annual Meeting of the Association for Computational Linguistics*, pages 486–496, Sofia, Bulgaria.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

Ronan LeNagard and Philipp Kohen. 2010. Aiding Pronoun Translation with Co-Reference Resolution. In *The Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppasala, Sweden.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Assessing the Discourse Factors that Influence the Quality of Machine Translation. In *The 52nd Annual Meeting of the Association for Computational Linguistics*, pages 283–288, Baltimore, MD.

Chin-Yew Lin and Franz J. Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest

Common Subsequence and Skip-Bigram Statics. In *The 42nd Meeting of the Association for Computational Linguistics*, pages 605–612, Barcelona, Spain.

Annie Louis and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistics*, 39(2):267–300, June.

Willian C. Mann and Sandra A. Thompson. 1987. *Rhetorical Structure Theory: A Theory of Text Organization*. Cambridge University Press, Cambridge, UK.

Daniel Marcu, Lynn Carlson, and Maki Watanabe. 2000. The automatic translation of discourse structures. In *The 1st North American chapter of the Association for Computational Linguistics conference*, pages 9–17. Association for Computational Linguistics, April.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using Sense-labeled Discourse Connectives for Statistical Machine Translation. In *The Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 129–138, Avignon, France.

Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. 2012. Machine translation of labeled discourse connectives. In *The Tenth Biennial Conference of the Association for Machine Translation in the Americas*, San Diego, CA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *The 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Daniele Pighin and Lluís Màrquez. 2011. Automatic Projection of Semantic Structures: an Application to Pairwise Translation R anking. In *The Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 1–9, Portland, OR.

Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *The Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Waikiki, Honolulu, Hawaii.

Andrei Popescu-Belis, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni, and Sandrine Zufferey. 2012. Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns. In *The Eighth International Conference on Language Resources and Evaluation*, pages 2716–2720, Istanbul, Turkey.

Marion Potet, Emmanuelle Esperança-Rodier, Laurent Besacier, and Hervé Blanchon. 2012. Collection of a Large Database of French-English SMT Output Corrections. In *The 8th International Conference on Language Resources and Evaluation*, pages 23–25, Istanbul, Turkey.

Raphael Rubino, Jos G. C. de Souza, Jennifer Foster, and Lucia Specia. 2013. Topic Models for Translation Quality Estimation for Gisting Purposes. In *The XIV Machine Translation Summit*, pages 295–302, Nice, France.

Carolina Scarton and Lucia Specia. 2014. Document-level translation quality estimation: exploring discourse and pseudo-references. In *The 17th Annual Conference of the European Association for Machine Translation*, pages 101–108, Dubrovnik, Croatia.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *The Seventh Biennial Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA.

Radu Soricut and Abdessamad Echihabi. 2010. TrustRank: Inducing Trust in Automatic Translations via Ranking. In *The 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - A translation quality estimation framework. In *The 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria.

Amalia Todirascu, Thomas Françoí, Nuria Gala, Cédric Fairon, Anne-Laure Ligozat, and Delphine Bernhard. 2013. Coherence and Cohesion for the Assessment of Text Readability. In *The 10th International Workshop on Natural Language Processing and Cognitive Science*, pages 11–19, Marseille, France.

Ferhan Ture, Douglas W. Oard, and Philip Resnik. 2012. Encouraging Consistent Translation Choices. In *The 12th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 417–426, Montréal, Quebec, Canada.

Billy T. M. Wong and Chunyu Kit. 2012. Extending Machine Translation Evaluation Metrics with Lexical Cohesion to Document Level. In *The 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level Consistency Verification in Machine Translation. In *The XII Machine Translation Summit*, pages 131–138, Xiamen, China.

Gong Zhengxian, Zhang Yu, and Zhou Guodong. 2010. Statistical Machine Translation Based on LDA. In *The 4th International Universal Communication Symposium*, pages 279–283, Beijing, China.