Searching for Context: a Study on Document-Level Labels for Translation Quality Estimation

Carolina Scarton[1], **Marcos Zampieri**[2,3], **Mihaela Vela**[2], **Josef van Genabith**[2,3] **and Lucia Specia**[1]
[1]University of Sheffield / Regent Court, 211 Portobello, Sheffield, UK
[2]Saarland University / Campus A2.2, Saarbrücken, Germany
[3]German Research Centre for Artificial Intelligence / Saarbrücken, Germany
{c.scarton,l.specia}@sheffield.ac.uk
{marcos.zampieri,m.vela}@uni-saarland.de
josef.van_genabith@dfki.de

Abstract

In this paper we analyse the use of popular automatic machine translation evaluation metrics to provide labels for quality estimation at document and paragraph levels. We highlight crucial limitations of such metrics for this task, mainly the fact that they disregard the discourse structure of the texts. To better understand these limitations, we designed experiments with human annotators and proposed a way of quantifying differences in translation quality that can only be observed when sentences are judged in the context of entire documents or paragraphs. Our results indicate that the use of context can lead to more informative labels for quality annotation beyond sentence level.

1 Introduction

Quality estimation (QE) of machine translation (MT) (Blatz et al., 2004; Specia et al., 2009) is an area that focuses on predicting the quality of new, unseen machine translation data without relying on human references. This is done by training models using features extracted from source and target texts and, when available, from the MT system, along with a quality label for each instance.

Most current work on QE is done at the sentence level. A popular application of sentence-level QE is to support post-editing of MT (He et al., 2010). As quality labels, Likert scores have been used for post-editing effort, as well as post-editing time and edit distance between the MT output and the final version – HTER (Snover et al., 2006).

There are, however, scenarios where quality prediction beyond sentence level is needed, most notably in cases when automatic translations without post-editing are required. This is the case, for example, of quality prediction for an entire product review translation in order to decide whether or not it can be published as is, so that customers speaking other languages can understand it.

The quality of a document is often seen as some form of aggregation of the quality of its sentences. We claim, however, that document-level quality assessment should consider more information than sentence-level quality. This includes, for example, the topic and structure of the document and the relationship between its sentences. While certain sentences are considered perfect in isolation, their combination in context may lead to incoherent text. Conversely, while a sentence can be considered poor in isolation, when put in context, it may benefit from information in surrounding sentences, leading to a document that is fit for purpose.

Document-level quality prediction is a rather understudied problem. Recent work has looked into document-level prediction (Scarton and Specia, 2014; Soricut and Echihabi, 2010) using automatic metrics such as BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) as quality labels. However, their results highlighted issues with these metrics for the task at hand: the evaluation of the scores predicted in terms of mean error was inconclusive. In most cases, the prediction model only slightly improves over a simple baseline where the average BLEU or TER score of the training documents is assigned to all test documents.

Other studies have considered document-level information in order to improve, analyse or au-

tomatically evaluate MT output (not for QE purposes). Carpuat and Simard (2012) report that MT output is overall consistent in its lexical choices, nearly as consistent as manually translated texts. Meyer and Webber (2013) and Li et al. (2014) show that the translation of connectives differs from humans to MT, and that the presence of explicit connectives correlates with higher HTER values. Guzmán et al. (2014) explore rhetorical structure (RST) trees (Mann and Thompson, 1987) for automatic evaluation of MT into English, outperforming traditional metrics at system-level evaluation.

Thus far, no previous work has investigated ways to provide a global quality score for an entire document that takes into account document structure, without access to reference translations. Previous work on document-level QE use automatic evaluation metrics as quality labels that do not consider document-level structures and are developed for inter-system rather than intra-system evaluation. Also, previous work on evalution of MT does not focus on complete evaluation at document-level.

In this paper, we show that the use of BLEU and other automatic metrics as quality labels do not help to successfully distinguish different quality levels. We discuss the role of document-wide information for document-level quality estimation and present two experiments with human annotators.

In the first experiment, translators are asked to subjectively assess paragraphs in terms of cohesion and coherence (herein, SUBJ). In the second experiment, a two-pass post-editing experiment is performed in order to measure the difference between corrections made with and without wider contexts (the tow passes are called PE1 and PE2, repectively).

The task of assessing paragraphs according to cohesion and coherence is highly subjective and thus the results of the first study did not show high agreement among annotators. The results of the two-stage post-editing experiment showed significant differences from the post-editing of sentences without context to the second stage where sentences were further corrected in context. This is an indication that certain translation issues can only be solved by relying on wider contexts, which is a crucial information for document-level QE. A manual analysis was conducted to evaluate differences between PE1 and PE2. Although several of the changes were found to be related to style or other non-discourse related phenomena, many discourse related changes were performed that were only possible given the wider context available.

In the remainder of this paper we first present related work in Section 2. In Section 3 we discuss the use of BLEU-style metrics for QE at document level. Section 4 describes the experimental set up used in the paper. Section 5 presents the first study were the annotators assess quality in terms of cohesion and coherence, while Section 6 shows the two-pass post-editing experiment and its results. The conclusions and future work are presented in Section 7.

2 Related work

The research reported here is about quality estimation at document-level. Therefore, work on document-level features and document-level quality prediction are both relevant, as well as studies on how discourse phenomena manifest in the output of MT systems.

Soricut and Echihabi (2010) propose document-level features to predict document-level quality for ranking purposes, having BLEU as quality label. While promising results were reported for ranking of translations for different source documents, the results for predicting absolute scores proved inconclusive. For two out of four domains, the prediction model only slightly improves over a baseline where the average BLEU score of the training documents is assigned to all test documents. In other words, most documents have similar BLEU scores, and therefore the training mean is a hard baseline to beat.

Scarton and Specia (2014) propose a number of discourse-informed features in order to predict BLEU and TER at document level. They also found the use of these metrics as quality labels problematic: the error scores of several QE models were very close to that obtained by the training mean baseline. Even when mixing translations from different MT systems, BLEU and TER were not found to be discriminative enough.

Carpuat and Simard (2012) provide a detailed evaluation of lexical consistency in translations of documents produced by a statistical MT (SMT) system, i.e., on the consistency of words and phrases in the translation of a given source text. SMT was found to be overall consistent in its lexi-

cal choices, nearly as consistent as manually translated texts.

Meyer and Webber (2013) present a study on implicit discourse connectives in translation. The phenomenon is evaluated using human references and machine translations for English-French and English-German. They found that humans translated explicit connectives in the source (English) into implicit connectives in the target (German and French) in 18% of the cases. MT systems translated explicit connectives into implicit ones less often.

Li et al. (2014) study connectives in order to improve MT for Chinese-English and Arabic-English. They show that the presence of explicit connectives correlates with high HTER for Chinese-English only. Chinese-English also showed correlation between ambiguous connectives and higher HTER. When comparing the presence of discourse connectives in translations and post-editions, they found that cases of connectives only appearing in the translation or post-edition also show correlation with high HTER scores.

Guzmán et al. (2014) explore RST trees (Mann and Thompson, 1987) for automatic evaluation of MT into English, with a discourse parser to annotate RST trees at sentence level in English. They compare the discourse units of machine translations with those in the references by using tree kernels to compute the number of common subtrees between the two trees. This metric outperformed others at system-level evaluation.

In summary, no previous work has investigated ways to provide a global quality score for an entire document that takes into account document structure, neither for evaluation nor for estimation purposes.

3 Automatic evaluation metrics as quality labels for document-level QE

As discussed in Section 2, although the use of BLEU-style metrics as quality scores for document-level QE clearly seems inadequate, previous work resorted to these automatic metrics because of the lack of better labels. In order to better understand this problem, we conducted an experiment with French-English translations from the LIG corpus (Potet et al., 2012). We took the first part of the corpus containing 119 source documents on the news domain (from various WMT news test sets), their MT by a phrase-based SMT system, a post-edited version of these translations by a human translator, and a reference translation. We used a range of automatic metrics such as BLEU, TER, METEOR-ex (exact match) and METEOR-st (stem match), which are based on a comparison between machine translations and human references, and the "human-targeted" version of BLEU and TER, where machine translations are compared against their post-editions: HBLEU and HTER. Table 1 shows the results of the average score (AVG) for each metric considering all documents, as well as the standard deviation (STDEV).

	AVG	STDEV
BLEU (↑)	0.27	0.05
TER (↓)	0.53	0.07
METEOR-ex (↑)	0.29	0.03
METEOR-st (↑)	0.30	0.03
HTER (↓)	0.21	0.03
HBLEU (↑)	0.64	0.05

Table 1: Average metric scores in the LIG corpus.

We conducted a similar analysis on the English-German (EN-DE) news test set from WMT13 (Bojar et al., 2013), which contains 52 documents, both at document and paragraph levels. Three MT systems were considered in this analysis: **UEDIN** (an SMT system), **PROMT** (a hybrid system) and **RBMT-1** (a rule-based system). Average metric scores are shown in Table 2.

For all the metrics and corpora, the STDEV values for documents are very small (below 0.1), indicating that all documents are considered similar in terms of quality according to these metrics (the scores are all very close to the mean).

At paragraph level (Table 2), the scores variation increases, with BLEU showing the highest variation. However, the very high STDEV values for BLEU (very close to the actual average score for all documents) is most likely due to the fact that BLEU does not perform well for short segments such as a paragraph due to the n-gram sparsity at this level, as shown in Stanojević and Sima'an (2014).

Overall, it is important to emphasise that BLEU-style metrics were created to evaluate different MT systems based on the same input, as opposed to evaluating different outputs of a single MT system, as we do here. The experiments in Section 6 attempt to shed some light on alternative ways to accurately measure document-level quality, with an emphasis on designing a label for document-level quality prediction.

	UEDIN				PROMT				RBMT-1			
	Document		Paragraph		Document		Paragraph		Document		Paragraph	
	AVG	STDEV	AVG	STDEV	AVG	STDEV	AVG	STDEV	AVG	STDEV	AVG	STDEV
BLEU (↑)	0.2	0.048	0.2	0.16	0.19	0.05	0.2	0.16	0.15	0.04	0.16	0.14
TER (↓)	0.62	0.063	0.63	0.24	0.61	0.07	0.62	0.25	0.66	0.06	0.67	0.23
METEOR-ex (↑)	0.37	0.056	0.37	0.16	0.36	0.06	0.37	0.16	0.32	0.05	0.33	0.15
METEOR-st (↑)	0.39	0.058	0.39	0.16	0.38	0.06	0.39	0.16	0.34	0.05	0.35	0.15

Table 2: Average metric scores for automatic metrics in the WMT13 EN-DE corpus.

4 Experimental settings

In the following experiments, we consider a paragraph as a "document". This decision was made to make the annotation feasible, given the time and resources available. Although the datasets are different for the two subtasks, they were taken from the same larger corpus and annotated by the the same group of translators.

4.1 Methods

The SUBJ experiment (Section 5) consists in assessing the quality of paragraphs in terms of cohesion and coherence. We define cohesion as the linguistic marks (cohesive devices) that connect clauses, sentences or paragraphs together; coherence captures whether clauses, sentences or paragraphs are connected in a logical way, i.e. whether they make sense together (Stede, 2011). In order to assess these two phenomena, we propose a 4-point scale. For coherence: 1=Completely coherent; 2=Mostly coherent; 3=Little coherent, and 4=Incoherent; for cohesion: 1=Flawless; 2=Good; 3=Disfluent and 4=Incomprehensible.

PE1 and PE2 (Section 6) consist in objective assessments through the post-editing of MT sentences in two rounds: in isolation and in context. In the first round (PE1), annotators were asked to post-edit sentences which were shown to them out of context. In the second round (PE2), they were asked to further post-edit the same sentences now given in context and fix any other issues that could only be solved by relying on information beyond individual sentences. For this, each annotator was given as input the output of their PE1, i.e. the sentences they had previously post-edited themselves.

4.2 Data

The datasets were extracted from the test set of the EN-DE WMT13 MT shared task. EN-DE was chosen given the availability of in-house annotators for this language pair. Outputs of the **UEDIN** SMT system were chosen as this was the best participating system for this language pair (Bojar et al., 2013). For the SUBJ experiment, paragraphs were randomly selected from the full corpus.

For PE1 and PE2, only source (English) paragraphs with 3-8 sentences were selected (filter S-NUMBER) to ensure that there is enough information beyond sentence-level to be evaluated and make the task feasible for the annotators. These paragraphs were further filtered to select those with cohesive devices. Cohesive devices are linguistic units that play a role in establishing cohesion between clauses, sentences or paragraphs (Halliday and Hasan, 1976). Pronouns and discourse connectives are examples of such devices. A list of pronouns and the connectives from Pitler and Nenkova (2009) was considered for that. Finally, paragraphs were ranked according to the number of cohesive devices they contain and the top 200 paragraphs were selected (filter C-DEV). Table 3 shows the statistics of the initial corpus and the resulting selection after each filter.

	Number of Paragraphs	Number of Cohesive devices
FULL CORPUS	1,215	6,488
S-NUMBER	394	3,329
C-DEV	200	2,338

Table 3: WMT13 English source corpus.

For the PE1 experiment, the paragraphs in C-DEV were randomised. Then, sets containing seven paragraphs each were created. For each set, the sentences of its paragraphs were also randomised in order to prevent annotators from having access to wider context when post-editing. The guidelines made it clear to annotators that the sentences they were given were not related, not necessarily part of the same document, and that therefore they should not try to find any relationships among them. For PE2, sentences were put together in their original paragraphs and presented to the annotators as a complete paragraph.

4.3 Annotators

The annotators for both experiments are students of "Translation Studies" courses (TS) in Saarland University, Saarbrücken, Germany. All students were familiar with concepts of MT and with post-editing tools. They were divided in two sets: (i) *Undergraduate students (B.A.)*, who are native speakers of German; and (ii) *Master students (M.A.)*, the majority of whom are native speakers of German. Non-native speakers have at least seven years of German language studies. B.A. and M.A. students have on average 10 years of English language studies. Only the B.A. group did the SUBJ experiment. PE1 and PE2 were done by all groups.

PE1 and PE2 were done using three CAT tools: PET (Aziz et al., 2012), Matecat (Federico et al., 2014) and memoQ.[1] These tools operate in very similar ways in terms of their post-editing functionalities, and therefore the use of multiple tools was only meant to make the experiment more interesting for students and did not affect the results. SUBJ was done without the help of tools.

5 Coherence/cohesion judgements

Our first attempt to access quality beyond sentence level was to explicitly guide annotators to consider discourse, where the notion of "discourse" covers various linguistic phenomena observed across discourse units. Discourse units can be clauses (intra-sentence), sentences or paragraphs.

Six sets with 17 paragraphs each were randomly selected from FULL CORPUS and given to 25 annotators from the B.A. group (each annotator evaluated one set). The task was to assess the paragraphs in terms of cohesion and coherence, using the scale given. The annotators could also rely on the source paragraphs. The agreement for the task in terms of Spearman's rank correlation and the number of students per set are presented in Table 4. The number of annotators per set is different because some of them did not complete the task.

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
Annotators	3	3	4	7	6	2
Coherence	0.07	0.05	0.16	0.16	0.28	0.58
Cohesion	0.38	0.43	0.28	0.09	0.38	0.12

Table 4: Spearman's correlation for the SUBJ task.

A low agreement in terms of Spearman's ρ rank correlation was found for both cohesion (ranging from 0.09 to 0.43) and coherence (ranging from 0.05 to 0.28, having 0.58 as an outlier) evaluations. Naturally, these concepts are very abstract, even for humans, offering substantial room for subjective interpretations. In addition, the existence of (often many) errors in the MT output can hinder the understanding of the text altogether, rendering judgements on any specific quality dimension difficult to make.

6 Quality assessment as a two-stage post-editing task

Using HTER, we measured the edit distance between the post-edited versions with and without context. The hypothesis is that differences between the two versions are likely to be corrections that could only be performed with information beyond sentence level.

For PE1, paragraphs from C-DEV set were divided in sets of seven and the sentences were randomised in order to prevent annotators from having access to context when post-editing. For PE2, sentences were put together in their original paragraphs and presented to annotators in context. A total of 112 paragraphs were evaluated in 16 different sets, but only sets where more than two annotators completed the task are presented here (SET1, SET2, SET7, SET9, SET14 and SET15).[2]

6.1 Task agreement

Table 5 shows the agreement for the PE1 and PE2 tasks using Spearman's ρ rank correlation. It was calculated by comparing the HTER values of PE1 against MT and PE2 against PE1. "Annotators" shows the number of annotators per set.

The HTER values of PE1 against PE2 are low, as expected, since the changes from PE1 to PE2 are only expected to reflect discourse related issues. In other words, no major changes were expected during the PE2 task. The correlation in HTER between PE1 and MT varies from 0.22 to 0.56, whereas the correlation in HTER between PE1 and PE2 varies between -0.14 and 0.39. The negative figures mean that the annotators strongly disagreed regarding the changes made from PE1 to PE2. This can be related to stylistic choices made by annotators, although further analysis is needed to study that (see Section 6.3).

[1] https://www.memoq.com/

[2] Sets with only two annotators are difficult to interpret.

	SET1	SET2	SET5	SET6	SET9	SET10	SET14	SET15	SET16
Annotators	3	3	3	4	4	3	3	3	3
PE1 x MT - HTER	0.63	0.57	0.22	0.32	0.28	0.18	0.30	0.24	0.18
PE1 x PE2 - HTER	0.05	0.07	0.05	0.03	0.10	0.06	0.09	0.07	0.05
PE1 x MT - Spearman	0.52	0.50	0.52	0.56	0.37	0.41	0.71	0.22	0.46
PE2 x PE1 - Spearman	0.38	0.39	−0.03	−0.14	0.25	0.15	0.14	0.18	−0.02

Table 5: HTER values for PE1 against MT and PE1 against PE2 and Spearman's rank correlation values for PE2 against PE1.

6.2 Issues beyond sentence level

The values for HTER among annotators in PE2 against PE1 were averaged in order to provide a better visualisation of changes made in the paragraphs from PE1 to PE2. Figure 1 shows the results for individual paragraphs in all sets. The majority of the paragraphs were edited in the second round of post-editions. This clearly indicates that information beyond sentence-level can be helpful to further improve the output of MT systems. Between 0 and 19% of the words have changed from PE1 to PE2 (on average 7% of the words changed).

An example of changes from PE1 to PE2 related to discourse phenomena is shown in Table 6. In this example, two changes are related to the use of information beyond sentence level. The first is related to the substitution of the sentence *"Das ist falsch"* - literal translation of *"This is wrong"* - by *"Das ist nicht gut"*, which fits better into the context. The other change is related to explicitation of information. The annotator decided to change from *"Hier ist diese Schicht ist dünn"* - literal translation of *"Here, this layer is thin"* - to *"Hier ist die Anzahl solcher Menschen gering"*, a translation that better fits the context of the paragraph *"Here, the number of such people is low"*.

6.3 Manual analysis

In order to better understand the changes made by the annotators from PE1 to PE2 and also better explain the negative values in Table 5, we manually inspected the post-edited data. This analysis was done by senior translators who were not involved in the actual post-editing experiments. They counted modifications performed and categorised them into three classes:

Discourse/context changes: changes related to discourse phenomena, which could only be made by having the entire paragraph text.

Stylistic changes: changes related to translator's stylistic or preferential choices. These changes can be associated with the paragraph context, although they are not strictly necessary under our post-editing guidelines.

Other changes: changes that could have been made without the paragraph context (PE1), but were only performed during PE2.

The results are shown in Table 7. Low agreement in the number of changes and the type of changes among annotators is found in most sets. Although annotators were asked not to make unnecessary changes (stylistic), some of them made changes of this type (especially annotators 2 and 3 from sets 5 and 6, respectively). These sets are also the ones that show negative values in Table 5. Since stylistic changes do not follow a pattern and are related to the background and preferences of the translator, the high number of this type of change for these sets can be the reason for the negative correlation figures. In the case of SET6, annotator 2 also performed several changes classified as "other changes". This may have also led to negative correlation values. However, the reasons behind the negative values in SET16 could include other phenomena, since overall the variation in the changes performed is low. Further analysis considering the quality of the post-edition needs to be done in order to better explain these results.

7 Conclusions

This paper focused on judgements of translation quality at document level with the aim to produce labels for QE datasets. We highlighted issues with the use of automatic evaluation metrics for the task, and proposed and experimented with two methods for collecting labels using human annotators.

Our pilot study for quality assessment of paragraphs in terms of coherence and cohesion proved a very subjective and difficult task. Definitions of cohesion and coherence are vague and the annotators' previous knowledge can play an important role during the annotation task.

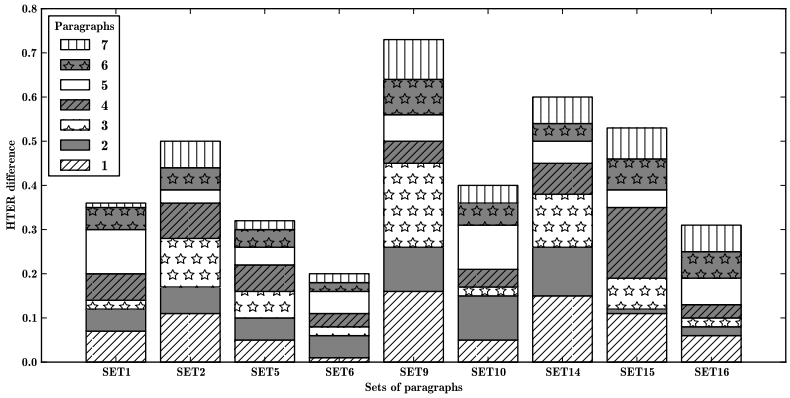

Figure 1: HTER between PE1 and PE2 for each of the seven paragraphs in each set.

PE1: - St. Petersburg bietet nicht viel kulturelles Angebot, Moskau hat viel mehr Kultur, es hat eine Grundlage. Es ist schwer fr die Kunst, sich in unserem Umfeld durchzusetzen. Wir brauchen das kulturelle Fundament, aber wir haben jetzt mehr Schriftsteller als Leser. **Das ist falsch**. In Europa gibt es viele neugierige Menschen, die auf Kunstausstellungen, Konzerte gehen. **Hier ist diese Schicht ist dünn**.
PE2: - St. Petersburg bietet nicht viel kulturelles Angebot, Moskau hat viel mehr Kultur, es hat eine Grundlage. Es ist schwer fr die Kunst, sich in unserem Umfeld durchzusetzen. Wir brauchen das kulturelle Fundament, aber wir haben jetzt mehr Schriftsteller als Leser. **Das ist nicht gut**. In Europa gibt es viele neugierige Menschen, die auf Kunstausstellungen, Konzerte gehen. **Hier ist die Anzahl solcher Menschen gering.**
SRC: - St. Petersburg is not a cultural capital, Moscow has much more culture, there is bedrock there. It's hard for art to grow on our rocks. We need cultural bedrock, but we now have more writers than readers. **This is wrong**. In Europe, there are many curious people, who go to art exhibits, concerts. **Here, this layer is thin.**

Table 6: Example of changes from PE1 to PE2.

	SET1			SET2			SET5			SET6				SET9				SET10			SET14			SET15			SET16		
Annotators	1	2	3	1	2	3	1	2	3	1	2	3	4	1	2	3	4	1	2	3	1	2	3	1	2	3	1	2	3
Discourse/context	2	3	1	0	6	2	2	1	0	2	2	0	0	1	7	1	0	4	0	0	1	0	1	2	1	2	0	1	1
Stylistic	2	0	1	1	0	1	3	11	0	0	3	9	3	5	10	1	3	1	2	2	6	0	0	3	3	2	2	1	3
Other	1	2	4	0	2	2	2	2	6	0	6	0	1	2	0	4	2	1	0	2	2	0	1	1	2	1	1	1	0
Total errors	5	5	6	1	8	5	7	14	6	2	11	9	4	8	17	6	5	6	2	4	9	0	2	6	6	5	3	3	4

Table 7: Manual analysis of PE1 and PE2.

Our second method for collecting labels using human annotators is based on post-editing and showed promising results on uncovering issues that rely on wider context to be identified (and fixed). Although some annotators did not follow the task specification and made unnecessary modifications or did not correct relevant errors at sentence level, overall the results showed that several issues could only be solved with paragraph-wide context. Moreover, even though stylistic changes can be considered unnecessary, some of them could only be made based on wider context.

We will now turn to studying how to use the information reflecting differences between the two rounds of post-editing as labels for QE at document level. One possibility is to use the HTER between the second and first rounds directly, but this can lead to many "0" labels, i.e. no edits made. Another idea is to devise a function that combines the HTER without context (PE1 x MT) and the difference between PE1 and PE2.

Our findings reveal important discourse dependencies in translation that go beyond QE, with relevance for MT evaluation and MT in general.

References

Aziz, Wilker, Sheila Castilho Monteiro de Sousa, and Lucia Specia. 2012. Cross-lingual Sentence Compression for Subtitles. In *The 16th Annual Conference of the European Association for Machine Translation*, pages 103–110, Trento, Italy.

Blatz, John, 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.

Bojar, Ondřej, 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.

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

Federico, Marcello, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, Alberto Massidda, Holger Schwenk, Loïc Barrault, Frederic Blain, Philipp Koehn, Christian Buck, and Ulrich Germann. 2014. THE MATECAT TOOL. In *The 25th International Conference on Computational Linguistics: System Demonstrations*, pages 129–132, Dublin, Ireland.

Guzmán, Francisco, 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.

Halliday, Michael A. K. and Ruqaiya Hasan. 1976. *Cohesion in English*. English Language Series. Longman, London, UK.

He, Yifan, Yanjun Ma, Josef van Genabith, and Andy Way. 2010. Bridging SMT and TM with Translation Recommendation. In *The 48th Annual Meeting of the Association for Computational Linguistics*, pages 622–630, Uppsala, Sweden.

Li, Junyi Jessy, 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.

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

Meyer, Thomas and Bonnie Webber. 2013. Implicitation of Discourse Connectives in (Machine) Translation. In *The Workshop on Discourse in Machine Translation*, pages 19–26, Sofia, Bulgaria.

Papineni, Kishore, 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.

Pitler, Emily and Ani Nenkova. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *The Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the Fourth International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 13–16, Suntec, Singapore.

Potet, Marion, 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.

Scarton, Carolina 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.

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 Seventh biennial conference of the Association for Machine Translation in the Americas*, AMTA 2006, pages 223–231, Cambridge, MA.

Soricut, Radu 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.

Specia, Lucia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *The 13th Annual Conference of the European Association for Machine Translation*, pages 28–37, Barcelona, Spain.

Stanojević, Miloš and Khalil Sima'an. 2014. Fitting Sentence Level Translation Evaluation with Many Dense Features. In *2014 Conference on Empirical Methods in Natural Language Processing*, pages 202–206, Doha, Qatar.

Stede, Manfred. 2011. *Discourse Processing*, volume 4 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.