A Reading Comprehension Corpus for Machine Translation Evaluation

Carolina Scarton and Lucia Specia
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
{c.scarton, l.specia}@sheffield.ac.uk

Abstract

Effectively assessing Natural Language Processing output tasks is a challenge for research in the area. In the case of Machine Translation (MT), automatic metrics are usually preferred over human evaluation, given time and budget constraints. However, traditional automatic metrics (such as BLEU) are not reliable for absolute quality assessment of documents, often producing similar scores for documents translated by the same MT system. For scenarios where absolute labels are necessary for building models, such as document-level Quality Estimation, these metrics can not be fully trusted. In this paper, we introduce a corpus of reading comprehension tests based on machine translated documents, where we evaluate documents based on answers to questions by fluent speakers of the target language. We describe the process of creating such a resource, the experiment design and agreement between the test takers. Finally, we discuss ways to convert the reading comprehension test into document-level quality scores.

Keywords: Machine Translation, Reading Comprehension, Quality Estimation

1. Introduction

Evaluating Machine Translation (MT) systems outputs is a challenging task. Whether the evaluation goal is to compare MT systems, to inform end-users or to assist in the translation process (such as in post-editing), appropriate evaluation methods and metrics need to be applied in order to provide reliable assessments.

Automatic metrics that contrast system outputs against reference translations, such as BLEU (Papineni et al., 2002), are widely explored to compare MT systems and measure the progress of a given MT system over time. **Quality Estimation (QE)** is a different evaluation method that provides a quality prediction for new, unseen machine translated texts, without relying on reference translations (Blatz et al., 2004; Specia et al., 2009). These metrics are useful to inform end-users and post-editors. Most work in QE focuses on sentence-level and word-level prediction (Callison-Burch et al., 2012; Bojar et al., 2013; Bojar et al., 2014; Bojar et al., 2015). Estimation of quality at sentence and word levels are probably the most useful types of prediction for post-editing, since post-editors can benefit from smaller parts of the document that are already acceptable, instead of relying on a single quality score for the entire document. On the other hand, **document-level QE** can be desirable for applications aimed at other types of end-users (such as *gisting*) and where fully automated MT is needed (e.g. because the amount of data is unfeasible for human post-editing).

While for sentence and word-level QE several quality labels have been proposed so far (e.g. HTER (Snover et al., 2006), *likert*), there is a lack of studies in quality labels for document-level. Previous work use BLEU-style metrics as labels (Scarton and Specia, 2014; Soricut and Echihabi, 2010; Scarton, 2015) . The WMT15 QE shared task also followed this approach, using METEOR (Banerjee and Lavie, 2005) for a paragraph-level QE task (Bojar et al., 2015). However, as shown by Scarton et al. (2015), these metrics do not distinguish well among documents, i.e. most documents produced by the same or a similar MT system show similar quality scores.

One issue of document-level quality labeling, noted by Scarton et al. (2015), is that the task of asking humans to assess documents is not trivial. While *likert* scores can be successfully applied for sentence and word levels, document-level quality can not be evaluated in the same way. Issues at document level should include problems at all other levels (word, sentence), plus problems at discourse level. Isolating these different types of problems from one another can provide a solution, but a costly one. The main question – "what is the quality of a document?" – thus remain unanswered.

In this paper we present a reading comprehension corpus that has been machine translated by different MT systems and by a human translator, with the aim of developing a new quality label for entire translated documents. Our hypothesis is that if readers of the target language can answer (manually written) reading comprehension questions on the texts accurately, the document translation is a good translation. Conversely, the translation is of bad quality. This corpus is an extension of the CREG corpus[1] (Ott et al., 2012) by taking the original documents in German and translating them (and the questions and answers) into English. Questions about each document were answered by paid volunteers, fluent English speakers, through Google Forms[2] questionnaires. Volunteers were staff members and students of the University of Sheffield, UK.

Section 2. introduces related work. Section 3. presents the CREG corpus and the pre-processing steps applied to it. Section 4. shows the experimental settings, how the data was collected and the agreement between test takers.

[1] http://www.uni-tuebingen.de/en/research/core-research/collaborative-research-centers/sfb-833/section-a-context/a4-meurers.html

[2] https://www.google.co.uk/forms/about/

2. Related Work

The usefulness of reading comprehension tests for MT evaluation has been addressed in previous work. Jones et al. (2005a) use the Defence Language Proficiency Test (DLPT) structure to evaluate the readability of Arabic-English MT texts. Their results show that subjects are slower at answering questions on the machine translated documents and that their accuracy is also inferior compared to human translated documents. Jones et al. (2005b) also use the DLPT-style questions, aiming to find which level of Arabic reading comprehension a native speaker of English could achieve by reading a machine translated document. Their results show that MT texts led to an intermediate level of performance by English native speakers. They also show that, in terms of BLEU, the performance of documents in different levels do not degrade as indicated by the reading comprehension evaluation. This shows that BLEU is clearly not adequate to distinguish between different documents machine translated by the same MT system.

Berka et al. (2011) use a quiz-based approach for MT evaluation. They collected a set of texts in English, created yes/no questions in Czech about these texts and machine translated the English texts by using four different MT systems. The texts consist of small paragraphs (one to three sentences) from various domains (news, directions descriptions, meeting and quizes). Their results show that outputs produced by different MT systems led to different accuracy in the annotators' answers.

Our work differs from previous because our focus is on the general quality of machine translated documents. Whilst Jones et al. (2005a) and Jones et al. (2005b) focus on levels of literacy and Berka et al. (2011) address MT system comparison, our research is more general. More specifically, our aim is to investigate ways to go from answers to questions based on machine translated documents to a document-level quality label that encompasses, in an abstract way, document-wide translation issues. With this quality label, we expect to be able to distinguish machine translated documents, one from another, independent on the MT systems that produced them. Our work also differs from Berka et al. (2011) because our documents are larger: they used documents with 1 to 3 sentences, while we use documents with average length of 46.95 sentences.

3. Corpus

CREG is a corpus of German documents with reading comprehension questions created for the purpose of assessing the proficiency level of second language learners. It has over 100 original documents of various genres (e.g. literature, news). The reading comprehension exercises (questions) expect open, descriptive answers. Although one can argue that multiple choice questions are straightforward to correct and probably easier to convert into quality scores, they can bias the answers of the test takers (they can more easily try to guess the answers). Each question has one or more gold standard answers. In our studies, we consider the following releases: CREG-5K (96 documents) and CREG-TUE (21 documents). Together, these releases contain a large number of distinct documents (117) and questions/answers, for which all actual documents are available, and not just the questions.

A sequence of pre-processing steps was applied to this corpus to prepare it for our experiments. Firstly, since sentence boundaries are important for Statistical Machine Translation (SMT) systems, one of the translation approaches considered in this paper, we corrected the corpus to use appropriate hard returns for sentences, according to punctuation (the original XML files did not encode this information). Secondly, some documents were repeated, each copy with different questions. For these cases, the questions were merged and only one copy was considered. Thirdly, some documents were a combination of two or more stories. In order to make the evaluation process more feasible (with shorter documents for test takers), we split these documents, considering each story a different document. Additionally, some documents were too long, which would probably make takers give up on answering the questionnaires. Therefore, we removed parts of such documents that were not important for answering the questions and did not affect document coherence. Finally, documents with less than three questions were discarded to avoid having documents with questions that were too easy to answer. The statistics of our corpus are given in Table 1. CREG refers to the original version, while CREG-clean, to the preprocessed version.

CREG-clean was then machine translated by different MT German-English systems: **Google Translate**[3], **Bing Translator**[4], **SYSTRAN**[5] and a **MOSES** baseline system[6]. The MOSES system was trained with WMT15 data (Bojar et al., 2015). These systems correspond to the state-of-the-art in SMT and a Rule-based Machine Translation (RBMT). In order to generate examples potentially having more problems in terms of cohesion and coherence (document-level problems), we also generated a version of each document containing alternating sentences from each of the MT systems (referred to hereafter as **Mixed**). These resulted in five versions of the corpus.

As a control group to evaluate whether the questions can be answered given perfect translations (i.e. to make sure incorrect answer do not only stem from the fact that questions or documents are too complex), professional translations of a subset of the documents were also include in our data (36 documents). We refer to these as **oracle** translations. The questions and the gold-standard answers for each document were all translated by a professional translator. The machine translated corpus can be download from `https://github.com/carolscarton/CREG-MT-eval`.

An example of a document and its questions is given in Table 2. A machine translation (Google) and a human translation are also shown in the table. It is possible to observe that, given only the MT output and/or only English knowledge, it is difficult to answer correctly questions 2, 3 and 4.

In order to evaluate whether BLEU-style scores can distin-

[3] `https://translate.google.co.uk/`

[4] `https://www.bing.com/translator/`

[5] `http://www.systransoft.com/`

[6] `http://www.statmt.org/moses/?n=moses.baseline`

	words per document	sentences per document	documents	questions per document	words per question
CREG	673.91	65.08	117	8.83	11.94
CREG-clean	372.38	32.52	108	8.79	13.49

Table 1: CREG corpus statistics.

Original:
Objektbeschreibung einer 15-jährigen Wohneinheit Am Ende der Straße umgeben von Einfamilienhäusern erreichen Sie Ihr neues Heim. Von der Eingangstür treten Sie in den oberen Bereich dieser wunderbaren Wohneinheit, die die Eigentümer sehr sorgfältig und mit Liebe zum Detail removiert haben. Im Erdgeschoss befinden sich ein Bad mit Tageslicht, Gäste WC, die Küche und ein äußerst geräumiges Wohn/Esszimmer mit faszinierendem Blick auf den gepflegten Garten. Die Treppe hinunter sind ein weiteres Bad mit Dusche - bisher noch nicht benutzt - sowie zwei gleich große Räume, beide mit Blick auf den herrlichen Garten und das angrenzende Waldgebiet. Die Zimmer in diesem Bereich sind in hochwertigem Laminat ausgelegt. Wenn Sie verkehrsgünstig wohnen möchten und gleichzeitig eine familiäre Umgebung schätzen, ist diese Wohnung für Sie richtig.
Questions:
1- Für wen ist diese Wohnung ideal? 2- Ist die Wohnung in einem Neubau oder einem Altbau? 3- Nennen Sie zwei Zimmer im Erdgeschoss. 4- Wo ist die Wohnung? 5- Wie viele Zimmer gibt es im Keller?
MT (Google):
Description a 15-year residential unit At the end of the street surrounded by family houses you reach your new home. From the front door you enter into the upper region of this wonderful residential unit who remo four very carefully and with attention to detail the owners. Downstairs there is a bathroom with daylight, guest toilet, kitchen and an extremely spacious living / dining room with a fascinating view are the landscaped garden. The stairs are a further bathroom with shower - not yet used - and two equally sized rooms, both overlooking the beautiful garden and the adjacent forest. The rooms in this area are designed in high-quality laminate. If you want to stay conveniently and simultaneously appreciate a family environment, this apartment is right for you.
Questions:
1- For whom is this apartment ideal? 2- Is the apartment in a new building or an old building? 3- Name two rooms on the ground floor. 4- Where is the apartment? 5- How many rooms are in the basement?
Human Translation:
Property description for a 15-year-old residential unit Your new home is at the end of the street surrounded by single-family homes. When you enter the front door, you find yourself on the upper floor of this wonderful property which the owners have carefully renovated and decorated with much attention to detail. The ground floor has a bathroom with natural light, a guest toilet, the kitchen and a spacious living/dining room with a fascinating view of the beautiful garden. Downstairs you will find an additional bathroom with shower (that has not yet been used) and two equally large bedrooms overlooking the wonderful garden. The downstairs rooms have high-quality laminate flooring. If you want to enjoy the benefits of a convenient location with a suburbian flair, this property is perfect for you.

Table 2: Example of a document in CREG corpus and its machine and human translations.

guish among documents of the same MT system or among them all, we calculated BLEU, TER (Snover et al., 2006) and METEOR scores for the 36 documents of CREG corpus that we have oracle translations. Table 3 shows these scores (all metrics were calculated by using the ASIYA toolkit (Giménez and Màrquez, 2010)).

All metrics show low standard deviation (STDEV) scores when calculated for the same system. Considering all systems, BLEU presents higher variation, which can be explained by the variation of systems (although, TER and METEOR did not show the same results for all systems together). Our hypothesis is that BLEU-style metrics will always follow the behaviour described in Scarton et al. (2015), showing similar prediction values for all documents translated by the same MT system.

	Moses		Google		Bing		Systran		Mixed		All	
	AVG	STDEV	AVG	STDEV	AVG	STDEV	AVG	STDEV	AVG	STDEV	AVG	STDEV
BLEU (↑)	0.259	0.078	0.313	0.071	0.280	0.084	0.120	0.062	0.245	0.073	0.243	0.100
TER (↓)	0.551	0.090	0.495	0.092	0.529	0.109	0.601	0.100	0.541	0.098	0.543	0.104
METEOR (↑)	0.303	0.046	0.340	0.044	0.319	0.058	0.207	0.043	0.290	0.047	0.292	0.066

Table 3: Average and standard deviation values for BLEU, TER and METEOR scores for the CREG corpus.

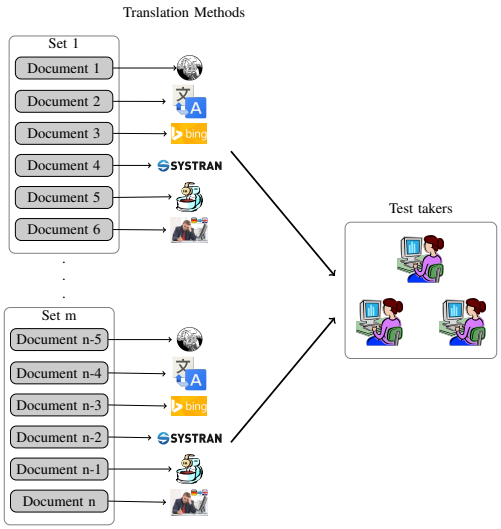

Figure 1: Split of corpus in sets and translation approach.

4. Data Acquisition and Evaluation

The corpus described in the previous section was divided in sets with six documents each.

The order of the documents was randomised before the splits in sets were created. Each set contains six different documents, which were also translated by different means. Figure 1 shows the structure of the sets and how the documents in each set were translated. The first document of each set was translated by Moses, the second by Google, the third by Bing, the forth by Systran, the fifth by the mixed version with all systems and the sixth by a human translator (*Scenario 1*). In a second scenario (*Scenario 2*), the order of the MT systems was different: Mixed, Systran, oracle, Moses, Bing and Google. Therefore, our experiments include two scenarios, varying the order of the MT systems that translated the documents.

In total, 19 different sets were created for each scenario. Each set was given to one volunteer test taker using an online questionnaire produced with Google Forms. The test taker was asked to answer questions in English.

The guidelines were similar to those used in reading comprehension tests: we asked the test takers to answer the questions using only the document provided. The original document (in German) was not given, therefore, test takers were not required to know German, but rather to speak fluent English. They were paid per questionnaire (set) and they were not able to evaluate the same set twice to prevent them from seeing the same document translated by a different system. Five sets were selected to be annotated five times, each time by a different test takers, so that agreement between test takers could be calculated.

4.1. Question classification

As previously mentioned, the reading comprehension questions are open questions, and thus any answer could be provided by the test takers. Another important detail is that these questions have different levels of complexity, meaning that some questions require more effort to be answered. Since our aim is to generate quality labels from the answers, information about the question complexity level is important. We therefore manually classified the questions using the classes in (Meurers et al., 2011), focusing on question forms and comprehension types (Day and Park, 2005).

Question forms: these can be directly defined by the question structure and by the expected answer. The question forms available in the CREG corpus are:

- **Yes/no questions:** are simple questions that admit either yes or no as valid answers.
- **Alternative questions:** are a combination of yes/no questions, connected with the connective "or".
- **True/false questions:** assume only true or false as a valid answer.
- **Wh-questions:** questions beginning with where, what, when, who, how, and why.

Comprehension types: in order to identify the type of comprehension that a question encode, one needs to read the text and identify the answer. The types of comprehension in questions in CREG are:

- **Literal questions:** can be answered directly from the text. They refer to explicit knowledge, such as facts, dates, location, names.
- **Reorganisation questions:** are also based on literal text understanding, but the test taker needs to combine information from different parts of the text to answer these questions.
- **Inference questions:** cannot be answered only with explicit information from the text and involve combining literal information with world knowledge.

4.2. Test takers agreement

The agreement was calculated by using Fleiss' Kappa metric. This metric is an extension of the Kappa metric allowing agreement calculations over more than two annotators. Alternatively, Spearman's ρ correlation coefficient was also calculated as the average between the ρ figure between each pair of test takers. Table 4 show results for Fleiss' Kappa and Spearman's ρ for the five sets.

	Scenario 1		Scenario 2	
	Fleiss' Kappa	ρ	Fleiss' Kappa	ρ
set 1	0.461	0.318	0.490	0.334
set 2	0.269	0.187	0.245	0.102
set 3	0.324	0.283	0.193	0.099
set 4	0.581	0.577	0.342	0.203
set 5	0.328	0.274	0.211	0.110

Table 4: Test takers agreement per set.

All sets except *set 3* from *Scenario 2* show 'fair' or 'moderate' agreement according to Fleiss' Kappa. Spearman's rho values are directly proportional to Fleiss. The best agreement is found in *set 4* from *Scenario 1* (0.581 for Fleiss' Kappa and 0.577 for Spearman's ρ) and the worse in *set 3* (0.269 and 0.187 for Fleiss' Kappa and Spearman's ρ, respectively).

We conducted further analysis on the data in an attempt to identify why some sets showed worse results than others. Firstly, we hypothesised that sets with lower agreement figures could contain more difficult questions, in other words, more questions classified as 'reorganisation' and 'inference'. However, this hypothesis proved false, since *set 3* (*Scenario 2*) only has literal questions and *set 4* (*Scenario 1*) has a mixed of all types of questions.

We also computed the correlation between the number of words in a set and its Fleiss' Kappa agreement. Table 5 shows the number of words and sentences per set. The correlation as calculated by Spearman's ρ was -0.60, indicating that when the number of words increases, the agreement decreases. However, it is worth noticing that *set 3* from *Scenario 2*, that showed the worst agreement, is not the largest set in terms of words.

	Scenario 1	Scenario 2
	Number of words	Number of words
set 1	2221	2230
set 2	3110	3152
set 3	2390	2391
set 4	2090	3937
set 5	2286	2343

Table 5: Number of words per set.

Table 6 shows Fleiss' Kappa values per document in all sets. Some documents show very low or no agreement, indicating that humans had problems answering questions for those documents. Although it would be expected that test takers should perform better when answering questions on human translated documents, such documents present low agreement in the majority of the sets (values in bold in Table 6).

Table 7 shows the average agreement per system, considering all machine translated documents (12 documents per system in total). MOSES is the system that showed highest agreement on average, followed by Bing. The worst agreement on average was found for Systran.

4.3. Question Marking

The most important component to generate our document-level quality label is the correctness of the questions. In order to mark the answers to the questions, we follow the work of (Ott et al., 2012), where the classes answers, based on the gold-standard (target) answer(s), are the following. For each of these classes, we assigned numeric scores (in brackets):

- **Correct answer:** the answer is a paraphrase of the target or an acceptable answer for the question (score = 1.0).
- **Extra concept:** incorrect extra concepts are added to the answer (score = 0.75).
- **Missing concept:** important concepts of the answer are missing (score = 0.5).
- **Blend:** mix of extra concepts and missing concepts (score = 0.25).
- **Non-answers:** the answer is completely incorrect (not related to the target answer) (score = 0.0.

We then manually marked all answers to all questions, by all test takers.

Devising a way to convert these "marks" into scores that corresponds to the overall quality of the document is a challenge. These scores need to take into account several peculiarities of the documents, such as document size, the number of questions and the complexity of the questions. Moreover, these scores need to represent fully correct answers and partially correct answers properly. This is the next step in our current work to be able to generate reliable quality scores that take into account document-level quality for MT *gisting* purposes.

	Fleiss' Kappa average
Moses	0.316
Google	0.221
Bing	0.300
Systran	0.167
Mixed	0.180
Human	0.211

Table 7: Average agreement per system.

5. Conclusions

In this paper we present the creation of a Reading Comprehension corpus for MT quality evaluation and estimation. This corpus is based on CREG, a corpus with German reading comprehension texts and exercises. Give these texts, we used different MT and human translations to generate sets of documents in English. The questions for each document were answered by fluent speakers of English.

We described the annotation process, the question type classification, and the question marking processes. Agreement among test takers was calculated and a discussion on its correlation to different phenomena provided. While it is not possible to draw conclusions about the reasons for low agreement in some sets of texts, our analysis addresses different aspects of the sets, such as number of words, type of questions and MT system.

Although previous work has addressed the use of reading comprehension questions for MT evaluation, our aims are

	Scenario 1					Scenario 2				
	set1	set2	set3	set4	set5	set1	set2	set3	set4	set5
doc 1	0.242	1.000	0.492	0.447	0.541	0.321	−0.071	0.048	0.333	−0.034
doc 2	0.301	0.275	0.200	0.207	0.327	0.363	0.176	0.021	0.476	−0.046
doc 3	0.644	0.528	0.253	0.254	0.182	**0.492**	**0.242**	**0.317**	**0.764**	**0.135**
doc 4	0.373	0.107	0.113	0.185	0.231	0.452	0.083	0.294	0.156	0.083
doc 5	0.321	−0.010	0.527	0.663	0.063	0.803	0.312	0.439	0.015	0.182
doc 6	**0.500**	**0.000**	**0.040**	**0.000**	**0.044**	0.417	0.299	0.044	−0.046	0.638

Table 6: Test takers Fleiss' Kappa agreement per document. It is worth noticing that document 1 (doc 1) in set 1 is different from doc 1 in set 2, doc 1 in set 3 and so on. Values in **bold** highlight values for human translation.

different. Whilst they focus either on levels of literacy or system comparison, the purpose of our corpus is to provide input for general MT evaluation. Our hypothesis is that reading comprehension questions can provide valuable information about the quality of an entire machine translated document. Future work include identifying the best way of use the question types, features of the texts, and correctness of answers as a quality score for documents.

6. Acknowledgements

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

7. Bibliographical References

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

Berka, J., Černý, M., and Bojar, O. (2011). Quiz-based evaluation of machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95:77–86.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence Estimation for Machine Translation. In *The 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *The Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

Bojar, O., Buck, C., Federman, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (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, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montreal, Canada.

Day, R. R. and Park, J.-S. (2005). Developing Reading Comprehension Questions. *Reading in a Foreign Language*, 17(1):60–73.

Giménez, J. and Màrquez, L. (2010). Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Jones, D. A., Gibson, E., Shen, W., Granoien, N., Herzog, M., Reynolds, D., and Weinstein, C. (2005a). Measuring Translation Quality by Testing English Speakers with a New Defense Language Proficiency Test for Arabic. In *International Conference on Intelligence Analysis*, McLean, VA.

Jones, D. A., Shen, W., Granoien, N., Herzog, M., and Weinstein, C. (2005b). Measuring Human Readability of Machine Generated Text: Studies in Speech Recognition and Machine Translation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA.

Meurers, R. Z., Ott, N., and Kopp, J. (2011). Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure. In *TextInfer 2011 Workshop on Textual Entailment*, pages 1–9, Edinburgh, UK.

Ott, N., Ziai, R., and Meurers, D. (2012). Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In T. Schmidt et al., editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Hamburg Studies on Multilingualism (Book 14), pages 47–69. John Benjamins Publishing Company, Amsterdam, The Netherlands.

Papineni, K., Roukos, S., Ward, T., and jing Zhu, W. (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.

Scarton, C. and Specia, L. (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.

Scarton, C., Zampieri, M., Vela, M., van Genabith, J., and Specia, L. (2015). Searching for Context: a Study on

Document-Level Labels for Translation Quality Estimation. In *The 18th Annual Conference of the European Association for Machine Translation*, pages 121–128, Antalya, Turkey.

Scarton, C. (2015). Discourse and document-level information for evaluating language output tasks. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 118–125, Denver, Colorado.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (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*, AMTA 2006, pages 223–231, Cambridge, MA.

Soricut, R. and Echihabi, A. (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, L., Turchi, M., Cancedda, N., Dymetman, M., and Cristianini, N. (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.