Exploring Adequacy Errors in Neural Machine Translation with the Help of Cross-Language Aligned Word Embeddings

Michael Ustaszewski
University of Innsbruck, Department of Translation Studies
michael.ustaszewski@uibk.ac.at

Abstract

Neural machine translation (NMT) was shown to produce more fluent output than phrase-based statistical (PBMT) and rule-based machine translation (RBMT). However, improved fluency makes it more difficult for post editors to identify and correct adequacy errors, because unlike RBMT and SMT, in NMT adequacy errors are frequently not anticipated by fluency errors. Omissions and additions of content in otherwise flawlessly fluent NMT output are the most prominent types of such adequacy errors, which can only be detected with reference to source texts. This contribution explores the degree of semantic similarity between source texts, NMT output and post edited output. In this way, computational semantic similarity scores (cosine similarity) are related to human quality judgments. The analyses are based on publicly available NMT post editing data annotated for errors in three language pairs (EN-DE, EN-LV, EN-HR) with the Multidimensional Quality Metrics (MQM). Methodologically, this contribution tests whether cross-language aligned word embeddings as the sole source of semantic information mirrors human error annotation.

1 Introduction

The most recent advances in artificial intelligence have brought substantial improvements to machine translation (MT). Systems based on artificial neural networks are able to produce more fluent and readable translations than most state-of-the-art phrase-based statistical (PBMT) and rule-based (RBMT) systems. The significant and highly promising advances notwithstanding, neural machine translation (NMT) still suffers from important shortcomings. Several lines of research address these shortcomings, most notably research on post-editing (PE) effort, on the evaluation and error annotation of NMT output and on (semi-)automated approaches to translation quality estimation.

Numerous studies in various language pairs and subject domains (see Section 2) have shown that NMT outperforms other types of MT in terms of fluency, while at the same time being more prone to adequacy errors such as omissions, additions or mistranslations. Adequacy errors are problematic from the perspective of the integration of NMT into translation workflows, because the identification and correction of adequacy errors is possible only by comparing NMT output to source segments, which arguably entails a higher cognitive load for post editors. Thus, participants in PE studies reported that NMT errors are more difficult to identify as compared other types of MT (Castilho et al. 2017). A phenomenon that is particularly difficult to handle for post editors and end users of NMT systems are *invisible* adequacy errors, first and foremost omissions in flawlessly fluent output that contains no traces of missing content, which means that they cannot be identified without the source text (van Brussel et al., 2018).

In view of these difficulties, the evaluation of semantic adequacy in NMT output and PE is indispensable to further advance the development of cutting-edge translation technology. Traditionally, the evaluation of MT output is performed by human annotators or post editors, but automated approaches have gained momentum as well (e.g. Moorkens et al., 2018; Specia et al., 2018). Semantic vector space models have become a cornerstone of present-day natural language processing (NLP) and as such, they play an important role in translation quality estimation, too. Cross-language embeddings trained in an unsupervised fashion (Artetxe, Labaka and Agirre, 2018; Joulin et al., 2018) are one of the most recent developments in distri-

Temnikova, I, Orăsan, C., Corpas Pastor, G., and Mitkov, R. (eds.) (2019) *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, Varna, Bulgaria, September 5 - 6, pages 122–128.
https://doi.org/10.26615/issn.2683-0078.2019_015

butional semantics, holding the potential to improve the performance of numerous multilingual NLP tasks.

Against this background, the present paper explores to what extent cross-language aligned word embeddings can be used to inform semantic analysis in NMT output evaluation. More specifically, the correspondence between human adequacy judgments and automatically generated semantic similarity scores is assessed. The main goal is to investigate whether publicly available, pre-trained cross-language embeddings as the sole source of semantic information (i.e. used in isolation without any other resources or features that capture the semantic relation between source and target segments) are reliable estimators of translation adequacy. The analyses are performed at the sentence level for three language pairs: English-German (EN-DE), English-Latvian (EN-LV), and English-Croatian (EN-HR), using publicly available error-annotated NMT and PE datasets.

2 Related Work

A number of error analysis studies have shown that NMT is prone to adequacy errors, i.e. deficiencies with regard to the semantic transfer of content from the source to the target language. Castilho et al. (2017) compared NMT to statistical MT and observed increases in fluency but at the same time there were more errors of omission, addition and mistranslation. For instance, in NMT omission errors accounted for 37% of all errors found in 100 Chinese-to-English translation segments from the patent domain, thus being the most frequent of seven error types, while for PBMT omission errors accounted only for 8% of all errors. Similar results were observed for four other language pairs in the domain of MOOC translations. Van Brussel et al. (2018) also observed numerous omission errors (13.1% of all adequacy errors) in a comparative evaluation of 665 English sentences translated by NMT, PBMT and RBMT into Dutch. The majority of omissions in NMT (85.5%) were due to missing content words, while for PBMT and RBMT these ratios were 70.0% and 0.1% respectively. As a consequence, most omission errors in NMT (69%) are invisible, i.e. not indicated by flawed fluency, whereas in the other two MT types, annotators deemed only 23% and 7% of omissions to be invisible without source text comparisons. The study concludes that due to their frequency and often invisible nature, adequacy errors are a major challenge to NMT and its users. Finally, Klubička, Toral and Sánchez-Cartagena make similar observations for the EN-HR pair, concluding that "NMT tends to sacrifice completeness of translation in order to increase overall fluency" (2018, 209). All these reviewed studies employ manual human error annotation to assess the quality of MT output. From a more technical perspective, Tu et al. (2016) argue that NMT's tendency to produce over- or under-translation is because conventional systems do not maintain a coverage vector.

A complementary line of research is concerned with the automated estimation of MT output quality at run-time without the use of reference translations (Specia et al., 2018). Translation quality estimation usually requires (a certain amount) of supervision and thus human-annotated training data. Given this interdependence of human and automated approaches to quality estimation, the present contribution sets out to relate automatically generated semantic similarity scores at the sentence level with human error annotation.

3 Materials and Methods

3.1 Datasets

The analyses are based on three publicly available datasets that provide fine-grained error annotation of NMT output according to the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014). For EN-DE and EN-LV, two datasets developed within the QT21 project (Specia et al., 2017) were used, each containing 1800 source sentences paired with the corresponding error-annotated NMT outputs and post-edited versions, 200 of which were annotated by two annotators. For EN-HR, the dataset by Klubička, Toral and Sánchez-Cartagena (2018) was used; it contains 100 source sentences together with error annotations of NMT output performed by two evaluators. Instead of post-edited target language versions, it contains human reference translations for 93 out of 100 source sentences.

From the original datasets, the raw text as well as error counts per sentence for each error type were extracted. Since the EN-HR dataset employs a customized, slightly extended version of the MQM error typology, the union of both typologies was used in this study. The two typologies are described in detail in Specia et al. (2017) and Klubička, Toral and Sánchez-Cartagena (2018).

For each dataset, only the annotations of the first evaluator were considered; however, to assess the quality of annotation, Cohen's kappa scores for inter-rater agreement on the annotation of omission errors were computed, indicating weak to moderate agreement. Summary statistics of the extracted data are shown in Table 1.

Pair	N	Tok	Errors			Kappa
			Tot	Flu	Acc	
EN-DE	1800	18.7	1.9	0.7	1.1	0.60
EN-LV	1800	22.2	1.9	0.9	1.0	0.52
EN-HR	93	20.5	1.4	0.8	0.4	0.39
Overall	3693	20.5	1.9	0.80	1.0	-

Table 1: Summary of datasets. Means are given for number of tokens and of total/fluency/adequacy errors per sentence. Cohen's kappa for agreement on omission annotation.

3.2 Cross-Language Aligned Word Embeddings

All sentences under investigation were represented as 300-dimensional word embedding vectors. To enable semantic analyses across source and target languages, pre-trained cross-language aligned *fastText*[1] word embeddings based on Wikipedia (Joulin et al., 2018) were used. In addition, for the EN-DE pair, custom cross-language aligned *fastText* embeddings we trained by aligning monolingual *fastText* Wikipedia embeddings[2] with the help of the *VecMap* toolkit[3] for cross-language word embedding mapping (Artetxe, Labaka and Agirre, 2018). For the mapping, the supervised mode of *VecMap* was used, based on the 5000-word EN-DE training dictionary from Artetxe, Labaka and Agirre (2018). Since both the pre-trained and custom embeddings are based on 300-dimensional *fastText* embeddings trained on Wikipedia, they are comparable irrespective of different mapping algorithms.

For each sentence in the dataset, the mean of the embeddings of each token in the sentence was calculated. The vector representations of the sentences in the datasets were built with the *flair* NLP library[4] implemented in Python (Akbik, Blythe and Vollgraf, 2018).

Subsequently, cosine similarity was computed between each source sentence and the following sentences:

(1) the corresponding NMT output;

(2) the corresponding PE target sentence (in the case of EN-DE and EN-LV) or the human reference translation (for EN-HR);

(3) a truncated copy of the NMT output, obtained by randomly removing 15% of its tokens;

(4) a truncated copy of the PE/reference translation sentence, obtained by randomly removing 15% of its tokens;

(5) two different sentences from the set of target sentences, randomly selected among the remaining target sentences in the given language (post-edited sentences for DE and LV, reference translation for HR);

(6) two different target language sentences, sampled from completely unrelated text collections: for DE and HR, sentences were sampled from the Universal Dependencies corpus (Nivre et al, 2016) included in the *flair* library, whereas for Latvian the *W2C* corpus[5] (Majliš and Žabokrtský, 2012) was taken as a source.

The inclusion of the sentences (3) to (6) was motivated by the need to test whether the combination of aligned word embeddings and cosine similarity adequately captures cross-linguistic similarity between sentences of varying degrees of semantic relatedness.

4 Results and Discussion

4.1 Similarity between Related vs. Unrelated Sentences

The comparison of similarity scores between source sentences and their machine-translated or post-edited equivalents on the one hand and randomly selected unrelated target sentences on the other provides insights into the general validity of the tested approach. The assumption is that the similarity between sentences in a translation relation – no matter whether machine-translated or post-edited – is higher than between unrelated pairs of source and target language sentences. What is more, it can be expected that among non-translated cross-lingual sentence pairs the similarity is higher when data is sampled from the same text collection,

[1] https://fasttext.cc/docs/en/aligned-vectors.html
[2] https://fasttext.cc/docs/en/pretrained-vectors.html
[3] https://github.com/artetxem/vecmap
[4] https://github.com/zalandoresearch/flair
[5] https://ufal.mff.cuni.cz/w2c

as opposed to data taken from a completely different corpus. Indeed, the results in Table 2 confirm this assumption, showing that the mean cosine similarity scores for translated source-target pairs (NMT and PE) are higher than for randomly aligned text pairs from the same dataset (MQM1, MQM2). The latter, in turn, are more similar to the source language sentences than sentence pairs obtained by assigning random sentences from unrelated corpora (X1, X2). The results are very similar for all three language pairs, suggesting the stability of cross-lingual word embeddings. It can also be seen that for EN-DE, the pre-trained and custom embeddings behave the same way, although the *VecMap*-aligned custom embeddings yield higher similarity scores. Both embedding types capture the differences between the semantically diverging sentences to the same extent.

Emb	NMT	PE	MQM1	MQM2	X1	X2
DE pre	0.22	0.22	0.16	0.16	0.07	0.07
DE cust	0.84	0.84	0.74	0.74	0.61	0.61
LV pre	0.30	0.31	0.25	0.25	0.17	0.17
HR pre	0.15	0.14	0.06	0.06	0.04	0.03

Table 3: Mean cosine similarity between source language sentences and the respective NMT and PE output, as well as randomly chosen target language sentences from the same corpus (MQM) and from different corpora (X). For DE, similarity scores were obtained from pre-trained (pre) and custom (cust) *fastText* embeddings.

4.2 Similarity between Closely-Related Sentences

To test whether the method is capable of detecting minor differences in meaning, NMT and PE outputs were juxtaposed with artificially truncated copies of these sentences by randomly removing 15% of tokens from the target sentences, not controlling for parts of speech, which means that punctuation may have been among the removed tokens. The truncation procedure is to simulate omissions in NMT output by creating semantically closely related sentences.

As shown in Table 3, the similarity scores between full vs. truncated sentences are almost identical, indicating that the method in isolation is not capable of capturing subtle semantic differences. Unlike the pre-trained embeddings, the *VecMap*-aligned embeddings do capture differences between full and truncated sentences, but the scores differ only marginally.

Emb		NMT	NMT_Short	PE	PE_Short
DE	pre	0.22	0.22	0.22	0.22
DE	cust	0.84	0.82	0.84	0.82
LV	pre	0.30	0.30	0.31	0.30
HR	pre	0.15	0.14	0.14	0.14

Table 2: Mean cosine similarity between source language sentences and the respective NMT and PE output, as well as copies of target sentences randomly truncated by 15% of tokens (NMT_Short, PE_Short). Scores provided for pre-trained (pre) and custom (cust) *fastText* embeddings.

As shown in Table 3, the similarity scores between full vs. truncated sentences are almost identical, indicating that the method in isolation is not capable of capturing subtle semantic differences. Unlike the pre-trained embeddings, the *VecMap*-aligned embeddings do capture differences between full and truncated sentences, but the scores differ only marginally.

It would be insightful to test whether truncations by more than 15% yield different results, and whether the removal of content words has a different impact on similarity and adequacy than the removal of function words or punctuation tokens. Preliminary exploration suggested that truncations by 30% do result in lower similarity scores, albeit only to a moderate extent. This might be due to the part-of-speech-insensitive nature of the employed truncation procedure, as well as to the use of context-insensitive word embeddings, as opposed to contextualized embeddings, such as *ELMo* (Peters et al., 2018), *BERT* (Devlin et al., 2019) or *flair* (Akbik, Blythe and Vollgraf, 2018) embeddings. Systematic analyses of the impact of truncation on similarity scores are left for future work.

The (almost) nonexistent differences between full and truncated sentences further suggest limitations as to the detection of omissions or additions as one of the most relevant types of NMT adequacy errors. Table 3 also shows that no tangible differences between NMT and PE were detected by either embedding type. This issue is discussed in more detail in the following subsection.

4.3 Correspondence between Cosine Similarity Scores and Human Error Annotation

Any valid computational metric should mirror human ratings, irrespective the fact that agreement between human raters is not always unanimous, especially in cognitively and intellectually demanding tasks. In the context of MT evaluation, it can be assumed that output containing adequacy errors, as assessed by human annotators or post-editors, exhibits lower degrees of semantic similarity according to vector space models. However, this observation was not made in this study.

Table 4 relates cosine similarity to the presence or absence of certain errors in NMT output: the first group compares machine-translated sentences that, according to human annotators, are free of adequacy errors (the left column of each block, designated with F for 'false') with sentences that contain at least one adequacy error (the second column of each block, designated with T for 'true'). The mean cosine scores for this group do not reveal any differences for NMT sentences that do and do not contain adequacy errors. Similar results were obtained for NMT output with and without omission errors (second group), for NMT output that does and does not contain only adequacy errors (third group), as well as for output that does and does not contain only fluency errors (fourth group). This lack of observed differences holds for both types of cross-language aligned embeddings used in the analyses, as shown in Table 4.

It was also tested whether the absence or presence of other error types and combinations thereof (e.g. output that contains mistranslations but no fluency errors) have an influence on cosine similarity scores, but no important differences were observed. In sum, the results clearly show that when used in isolation without any other resources or features, aligned cross-language word embeddings are hardly helpful to inform cross-linguistic similarity judgments in cases of subtle adequacy deviations typical of NMT.

	Adq Err		Omission		Only Adq		Only Flu	
Emb	**F**	**T**	**F**	**T**	**F**	**T**	**F**	**T**
DE pre	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22
DE cust	0.84	0.84	0.84	0.84	0.84	0.83	0.84	0.84
LV pre	0.30	0.31	0.30	0.31	0.31	0.30	0.30	0.30
HR cust	0.16	0.14	0.15	0.14	0.15	0.15	0.15	0.14

Table 4: Mean cosine similarity between source language sentences and the respective NMT output, grouped according to the absence and presence of four error types. Scores shown for pre-trained (pre) and custom (cust) *fastText* embeddings.

5 Conclusion

The measurement of cross-linguistic similarity is a highly complex problem with relevance not only to translation, but also, among other things, to semantic textual similarity (Agirre et al., 2016) or comparable and parallel corpus building (Sharoff, Rapp and Zweigenbaum, 2013). Recent advances in embeddings-based vector space representations have brought significant advances to cross-linguistic semantic problems, which can be useful in the context of translation quality estimation and MT evaluation.

The present study attempted to explore the usefulness of cross-language aligned word embeddings in isolation, i.e. without further resources or features. In doing so, the correspondence of cosine similarity scores has been related to human similarity judgments of NMT output and PE. It was observed that cross-language embeddings used in isolation are only able to differentiate between sentences related by translation on the one hand and unrelated in-domain and out-of-domain sentences on the other, which means that the analysis of subtle adequacy issues frequently observed in NMT, such as omissions or additions, requires more elaborate approaches. The results from the EN-DE language pair suggest that it makes no difference whether pre-trained *fastText* or custom *VecMap*-aligned cross-language embeddings are used, because both types do not capture subtle semantic differences. Analogous comparisons for other language pairs may yield more insights into the comparability of different types of cross-language word embeddings.

The methodology employed in this study could be improved in several ways. On the one hand, the embeddings in this study were used without any parameter tuning. On the other hand, contextualized word embeddings, such as *ELMo* (Peters et al., 2018), *BERT* (Devlin et al., 2019) or *flair* (Akbik, Blythe and Vollgraf, 2018), which were shown to yield state-of-the-art results in several NLP tasks, could be used as an alternative to the context-insensitive embeddings used in this study. However,

since the cross-language alignment of contextualized embeddings is a very recent and therefore still relatively unexplored line of research (e.g. Aldarmaki and Diab, 2019; Schuster et al., 2019), the use of contextualized cross-language aligned embeddings for the detection of subtle adequacy deviations is left for future work. A further potential improvement of the present methodology relates to the fact that in this study, sentences were represented as means of the embeddings of all words in the sentences. There are other approaches to compute sentence- or document-level embeddings from individual word embeddings (Chen, Ling and Zhu, 2018), and the *flair* library, for instance, implements various methods, such as minimum and maximum pooling or recurrent neural networks[6]. Similarly, there are alternatives to the traditionally used cosine similarity, for instance the word mover's distance (Kusner et al., 2015).

Given that monolingual embeddings are already being successfully employed in translation quality estimation (Specia et al., 2018), the unsupervised nature of cross-language embeddings may further promote this line of research. Yet, its application to translation quality estimation and error analysis requires more thorough benchmarking. This also means that human evaluation is still to be seen as pivotal to research into adequacy errors in NMT. Datasets that focus explicitly on omissions and additions might become an asset in this regard, since the datasets used in the present study are much wider in scope. While they do contain useful information about adequacy, complementary and more focused datasets might contribute to the development of new approaches to the automated detection of adequacy errors, including the problematic invisible omissions and additions.

Acknowledgments

I thank the four anonymous reviewers for their helpful comments.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In Proceedings of SemEval 2016. Association for Computational Linguistics, pages 497-511. http://dx.doi.org/10.18653/v1/S16-1081.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, pages 1638-1649. https://www.aclweb.org/anthology/C18-1139.

Hanan Aldarmaki and Mona Diab. 2019. Context-Aware Cross-Lingual Mapping. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pages 3906-3911. http://dx.doi.org/10.18653/v1/N19-1391.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 789–798. http://dx.doi.org/10.18653/v1/P18-1073.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is Neural Machine Translation the New State of the Art? In *The Prague Bulletin of Mathematical Linguistics*, 108, pages 109–120. https://doi.org/10.1515/pralin-2017-0013.

Qian Chen, Zhen-Hua Ling, and Xiaodan Zhu. 2018. Enhancing Sentence Embedding with Generalized Pooling. In *Proceedings of the 27th International Conference on Computational Linguistics*.Association for Computational Linguistics, pages 1815-1826. https://www.aclweb.org/anthology/C18-1154.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pages 4171-4186. http://dx.doi.org/10.18653/v1/N19-1423.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *Proceedings of the 2018*

[6] https://github.com/zalandoresearch/flair/blob/master/resources/docs/TUTORIAL_5_DOCUMENT_EMBEDDINGS.md

Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pages 2979-2984. http://dx.doi.org/10.18653/v1/D18-1330.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2018. Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian. In *Machine Translation*, 32, pages 195-2015. https://doi.org/10.1007/s10590-018-9214-x.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From Word Embeddings To Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015. JMLR: W&CP volume* 37.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. In *Revista Tradumàtica*, 12, pages 455-463. https://doi.org/10.5565/rev/tradumatica.77.

Martin Majliš and Zdeněk Žabokrtský. 2012. Language Richness of the Web. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Languages Resources Association (ELRA), pages 2927-2934. https://aclweb.org/anthology/papers/L/L12/L12-1110/.

Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty (eds.). 2019. *Translation Quality Assessment. From Principles to Practice*. Cham: Springer. https://doi.org/10.1007/978-3-319-91241-7.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), pages 1659-1666. https://www.aclweb.org/anthology/L16-1262.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, pages 2227-2237. http://dx.doi.org/10.18653/v1/N18-1202.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pages 1599-1613. http://dx.doi.org/10.18653/v1/N19-1162.

Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. 2013. Overviewing Important Aspects of the Last Twenty Years of Research in Comparable Corpora. In S. Sharoff, R- Rapp, P. Zweigenbaum and P. Fung (eds.). *Building and Using Comparable Corpora*. Berlin, Heidelberg: Springer, pages 1-17. https://doi.org/10.1007/978-3-642-20128-8_1.

Lucia Specia, Kim Harris, Frédéric Blain, Aljoscha Burchardt, Vivien Macketanz, Inguna Skadiņa, Matteo Negri, and Marcho Turchi. 2017. Translation Quality and Productivity: A Study on Rich Morphology Languages. In *Proceedings of the 16th Machine Translation Summit (Volume 1: Research Track),* pages 55- 71.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ram ón F. Astudillo, and André Martins. 2018. Findings of the WMT 2018 Shared Task on Quality Estimation. In *Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers*. Association for Computational Linguistics, pages 689–709. http://dx.doi.org/10.18653/v1/W18-6451.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling Coverage for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 76-85. http://dx.doi.org/10.18653/v1/P16-1008.

Laura Van Brussel, Arda Tezcan, and Lieve Macken. 2018. A Fine-grained Error Analysis of NMT, PBMT and RBMT Output for English-to-Dutch. In *Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan*. European Language Resources Association, pages 3799–3804. https://www.aclweb.org/anthology/L18-1600.