Automated Evaluation Metric for Terminology Consistency in MT

Kirill Semenov and **Ondřej Bojar**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské námestí 25, 118 00 Prague, Czech Republic
kir.semenow@yandex.ru
bojar@ufal.mff.cuni.cz

Abstract

The most widely used metrics for machine translation tackle sentence-level evaluation. However, at least for professional domains such as legal texts, it is crucial to measure the consistency of translation of terms throughout the whole text.

This paper introduces an automated metric for term consistency evaluation in machine translation (MT). To demonstrate the metric's performance, we used the Czech-to-English translated texts from the ELITR 2021 agreement corpus and the outputs of the MT systems that took part in WMT21 and WMT22 News Tasks. We show different modes of our evaluation algorithm and try to interpret the differences in the ranking of the translation systems based on standard sentence-level metrics and our approach. We also demonstrate that the proposed metric scores significantly differ from the widespread automated metric scores, and correlate with human assessment.

1 Introduction

Throughout the last decade, the quality of machine translation (MT) has improved significantly, and it is becoming a common phenomenon for various neural MT (NMT) systems to get better scores in manual direct assessment and other metrics than reference human translations (Akhbardeh et al., 2021; Bojar et al., 2018). However, such figures are obtained when the MT outputs are evaluated on the sentence level (i.e., each sentence is assessed separately, without context); in document-level evaluation, human translations typically remain the best, although exceptions exist (Popel et al., 2020). We can explain this situation by the fact that most of the current state-of-the-art NMT systems translate documents sentence by sentence, which thus can provoke inconsistencies in the translation of different linguistic elements – from anaphoric pronouns to named entities and terminology. We focus on the latter.

While term inconsistencies can be tolerable for the general spheres of communication, they are unacceptable for several professional domains, especially legal texts, where the coherent usage of terms is the ultimate characteristic.

In the case of the term translation in the legal domain, the goal of the MT system can be split into several parts:

1. To translate one source term to only one target term (we will call this property "consistency");
2. To ensure that every source term is mapped to a distinct target term (we will call this property "unambiguity");
3. To ensure that the target term is an adequate translation of the source term in general.

In this paper, we present a novel metric that focuses on the consistency and unambiguity of terms, whereas measuring the third parameter, adequacy, is delegated to the mainstream automated metrics such as BLEU (Papineni et al., 2002) and chrF (Popović, 2015). Our proposed metric can be applied automatically, and it needs a small amount of human preprocessing and annotation (for instance, it does not require reference translation of the sentences). However, it can include manually tuned parameters as a variable.

In Section 2, we describe the background in the field of the term consistency in MT; in Section 3 we introduce the algorithm of our metric; in Section 4 we present the data on which the metric is applied; in Section 5 we discuss the results and compare them to the widespread automated metrics in MT. Limitations of our method are highlighted in Section 6.

2 Background

Scholars have been drawing attention to document-level consistency for over a decade. For instance,

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 450 - 457
December 7-8, 2022 ©2022 Association for Computational Linguistics

Hardmeier (2012) presents a number of discourse-related phenomena (such as pronoun use and verb tense modeling) that should be taken into account, as well as an overview of the metrics that were designed to catch the consistency (by that moment, they did not correlate with human judgments much). Since this time, there have been various experiments in enhancing the sentence-level MT models for better consistency, by a variety of means, from hierarchical approaches (Ture et al., 2012) to post-editing the output sentences (Voita et al., 2019b). Notably, the main focus of the proposed systems tends to be on the discourse-related features of texts, such as verb forms, anaphora, ellipsis, named entities, etc. (Voita et al., 2019a), rather than on the terminology consistency.

There has also been progress in designing the evaluation for lexical consistency in domain-specific spheres. For example, the creators of SAO WMT test suite (Vojtěchová et al., 2019) point out that the most accurate evaluation for the audit reports is performed manually by professionals in the field, while neither the automated metrics nor the direct evaluation by non-experts gives valuable information about the ranking of the systems' quality. The same authors in 2020 introduced the concept of the "markables": the linguistic elements to which the human annotators have to pay special attention (Zouhar et al., 2020). In the paper, they considered the domains of Sublease, News, and Audit, and the main markables were the crucial terms in the document. The research reaffirms that the automated metrics such as BLEU are not very informative with respect to term consistency, while the non-professional annotators cannot spot the domain-specific inconsistencies; however, the additional annotation of the "markables" allows even the "lay" annotators to keep in mind the necessary terms, which makes the manual annotation more accurate and informative.

Another notable research by Alam et al. (2021) presents the ideas for automated metrics for the term consistency evaluation, namely, exact match accuracy, window overlap, and TER with bigger penalties for terms. The results of this approach, tested on the domain of medical texts about COVID-19, show a correlation with human professional judgments; however, for most of the metrics, reference translations or at least term dictionaries are necessary. Thus, the relevance of designing more automated metrics in the field is still valid.

3 Metric

Before explaining the metric in detail, we will reiterate our aim. Our first objective is to reward translation consistency (i.e., penalize the one-to-many correspondences in source-to-target term pairs). Secondly, we want translated terms to be unambiguous (i.e., we should penalize the many-to-one correspondences in source-to-target term pairs). Optionally, we also want to include adequacy in our estimation (i.e., penalize the inappropriate translation of the term); otherwise, we will rely on the widely accepted metrics for adequacy. Finally, we want the algorithm to be as automatic as possible, i.e., to avoid the necessity of human annotation on any level. To meet these demands, we introduce the following pipeline.

1. **General preprocessing**: We tokenize the texts. The tokenization needs to be consistent in both source and target texts to run the alignment algorithms (Step 3 below).

2. **Source terms extraction**: We extract "crucial" terms in the source text. The task can be reduced to keyword extraction, which has various approaches. In our study, we used the manual method based on regular expressions: in legal-like texts, the terms relevant for the document are announced uniformly at the beginning of the document (for example, by the phrases "hereinafter referred as..."). We justify this choice in Section 6. As a result of this step, we get a set of the terms that occur in the text (hereinafter: src term set), and, for each sentence, we get a list of terms that appear there.

3. **Term Alignment**: For automation, we suggest using any word alignment algorithm. In this experiment, we used fast-align algorithm introduced by Dyer et al. (2013). Now, for each text separately, we extract the alignments of the source terms obtained in Step 1.[1] At the end of this step, for each document, we have, firstly, lists of aligned target terms in each sentence, secondly, the dictionary of source terms and the counts of their corresponding alignments in this text (hereinafter: src-tgt dict).

[1] To create a better word alignment, we firstly collect all outputs of the same system into one text, apply fast-align to such big texts, and then split the alignments back to the initial document level.

4. **Choosing the "pseudo-reference" translations**: To measure the performance of the MT system, we have to compare the real occurrences of the translations (obtained in Step 3, hereinafter called "candidate" translations) to the translations that we expect to be used throughout the text (we call them "pseudo-reference" translations). Choosing the pseudo-reference translation is the trickiest element of the task. However, we can introduce several solutions to it. On the one hand, we can count the first occurrence of each translated term as the pseudo-reference. This is reasonable in the logic of legal texts, where the terms are "introduced" at the beginning and consistently used afterwards. On the other hand, we can choose the most frequent translation of the term to be the correct translation. In our experiment, we tried both approaches, which are easily done by the src-tgt dict or by the lists of the target terms for each sentence in the text. As a result of this step, we obtain the list of the "pseudo-reference" target terms for each sentence. Notably, the choice of the "pseudo-reference" terminology is calculated separately for each document.

5. **Evaluation**: After the four steps, the final data structure consists of quintuples, where each quintuple consists of the source sentence, the target sentence, and three lists: of the source terms, of the "candidate" occurrences of the translated source terms, and of the "pseudo-reference" translations. We can represent them as a variant of the TORT annotation (term-only reference translation, introduced by Bafna et al., 2021), where for each MT output sentence, there is a list of crucial reference terms instead of the whole text. Such lists of lists of "candidate" and "pseudo-reference" occurrences can be measured by the widespread data science metrics – multi-class precision, recall, true positive rate, etc. For better granularity, we also suggest grouping the lists by the source terms and counting the percentage of the correct occurrences of the exact term (hereinafter we call it "our" or "our own" metric).

Therefore, the main novelty of our approach is not the metric itself but an algorithm for automatizing the data collection for applying the widespread metrics.

4 Data

We used the data from the ELITR agreement test suite to test the metric. The test suite consists of various short agreement documents, namely, 18 purchase agreements, 13 lease and sublease agreements, and two agreements on renting or using the software. All documents have Czech as the source language and English as the target language; only for three files, the reference English translations are provided. As the MT outputs, we used the results of seven MT systems that took part in 2021 and 2022 competitions on this test suite. Detailed information about the systems is presented in Akhbardeh et al. (2021) and **?**, and the test suite texts are available online.[2]

5 Results and Discussion

In this section, we firstly comment on the absolute scores of the different variants of the proposed metric; secondly, we compare the ranking of the MT systems by our metric and by the ones represented in the findings of WMT21 and WMT22.

5.1 Proposed Metric Scores

Speaking about the absolute scores (see Table 1), we can see that for both years, if we fix formula that we use (either F1 or our own metric), the most frequent pseudo-reference initialization is regularly higher than the first-occurrence one (1-3% for F1; 3-5% for our metric). If we fix the pseudo-reference initialization and compare different formulas, the difference is bigger and varies between 7-9%. This can be a reflection of the fact that the NMT models are sentence based. The reason is following: if a model has a pre-trained distribution of translations for each term, then it may tend to choose the same likeliest translation for the term in the majority of the sentences. Thus, such likeliest translations will be most frequent in the src-tgt dicts, and will be chosen as "pseudo-references" in case of the most frequent initialization.

If we take into account the ranking of the algorithms, we can see that the big difference tends to be between the variants with different pseudo-reference choice. Kendall's tau paired comparisons between the variants support this hypothesis: the most correlating rankings are the F1 and our metric with first-occurrence initialization, next best correlation is between the F1 and our metric with the

[2] https://github.com/ELITR/agreement-corpus

Year	MT System	1st; F1	1st; Own	Freq; F1	Freq; Own	1st; F1 rank	1st; Own rank	Freq; F1 rank	Freq; Own rank
2021	**CUNI-Doc Transformer**	0.897	0.804	0.915	0.835	3	4	4	4
	CUNI-Transformer2018	0.857	0.776	0.895	0.827	8	7	8	7
	Facebook-AI	**0.907**	**0.838**	**0.930**	**0.871**	1	1	1	1
	Online-A	0.883	0.795	0.914	0.829	4	5	5	6
	Online-B	0.880	0.792	0.925	0.852	6	6	2	2
	Online-G	0.871	0.771	0.900	0.811	7	8	6	8
	Online-W	0.881	0.807	0.898	0.831	5	3	7	5
	Online-Y	0.900	0.813	0.921	0.840	2	2	3	3
2022	**ALMAnaCH-Inria**	0.816	0.688	0.885	0.807	11	11	10	9
	CUNI-Doc Transformer	0.897	0.805	0.916	0.836	4	6	4	6
	CUNI-Transformer	0.848	0.751	0.882	0.790	10	10	11	11
	JDExplore Academy	0.899	0.817	**0.928**	**0.863**	3	4	1	1
	Lan-Bridge	**0.902**	0.826	0.918	0.846	2	2	3	2
	Online-A	0.877	0.773	0.924	0.836	7	7	2	7
	Online-B	**0.902**	**0.831**	0.912	0.842	1	1	5	4
	Online-G	0.871	0.772	0.898	0.807	8	8	8	10
	Online-W	0.889	0.816	0.903	0.838	6	5	7	5
	Online-Y	0.860	0.767	0.892	0.809	9	9	9	8
	SHOPLINE-PL	0.895	0.822	0.910	0.845	5	3	6	3

Table 1: Scores of different metric variants. The first position in the column name denotes the method for choice of pseudo-reference ("Freq" for "most frequent translation", "1st" for "first occurrence"); the second means the metric ("F1" for F1 score and "Own" for our own metric – averaged percentage of the correct hits per term). The last four columns show the ranking of the systems.

Compared Setups	τ 2021	τ 2022
1st;F1 VS 1st;Own	.786*	.891*
1st;F1 VS Freq;F1	.643*	.636*
1st;F1 VS Freq;Own	.571	.673*
1st;Own VS Freq;F1	.429	.527*
1st;Own VS Freq;Own	.643*	.709*
Freq;F1 VS Freq;Own	.786*	.600*

Table 2: Pairwise Kendall's Tau correlations between the rankings of the scores obtained by different variants of our algorithm. The first column shows the pairs of variants we compare (separated by "VS"). The second and the third columns show Kendall's Tau scores; the asterisk denotes the values that are statistically significant for the null hypothesis of $\tau = 0 (p < 0.05)$.

same most frequent initialization. The next level of correlation is for the pairs of different initializations with the same metric (F1 or our own, respectively); the lowest correlation is between the most distant variants (such as F1 with the first-occurrence initialization and our metric with the most-frequent initialization). Notably, such a clear trend can be seen only on the results of WMT2021 systems, while on 2022 data, the only clear correlation is between the F1 and our metric with first occurrence initialization. The detailed tau values are shown in Table 2. Looking back at Table 1, we can see that, for 2021 systems, the best ones are Facebook-AI, Online-Y, and Online-B according to any metric variant, and the worst are CUNI-Transformer and Online-G. As for 2022 systems, the best-rated ones are JDExploreAcademy, Lan-Bridge, CUNI-DocTransformer, and Online-B, while the worst-rated ones are CUNI-Transformer ALMAnaCH-Inria, Online-Y, and Online-G.

5.2 Comparison with Standard Automatic Metrics and Direct Assessment

We also wanted to compare our metrics to the traditional manual and automated evaluation approaches for MT. Unfortunately, the only published results of the considered MT systems were based on the evaluation of another dataset of news texts, see Akhbardeh et al. (2021) and the actual scores online.[3] However, they can still give us an approximate idea of the systems' relative performance. For the 2021 news track, we have both automatic scores (BLEU and chrf) and human direct assessment, while for the 2022 track, we only have the automated metrics, the same as for the previous year. To compare the rankings of our metric and the standard ones, we find it logical to use Kendall's tau correlation, as it was applied in previous metrics shared tasks Macháček and Bojar (2014). The results of this comparison can be seen in Table 3. Regarding the WMT2021 outputs, on the one hand, the correlation between any automatic metric and any of our variants is not as high (and the p-values do not show any significance). The correlation with direct assessment scores, on the other hand, is high (more than 0.6 on average), and shows also the statistical significance in 2 out of 4 cases (for F1 with both variants of pseudo-reference initialization).

[3] https://github.com/wmt-conference/wmt22-news-systems

Metrics Compared	τ 2021	τ 2022
1st;F1 VS BLEU	.357	-.527
1st;F1 VS chrf	.286	
1st;F1 VS DA	.714*	N/A
1st;Own VS BLEU	.143	-.636
1st;Own VS chrf	.071	
1st;Own VS DA	.500	N/A
Freq;F1 VS BLEU	.143	-.527
Freq;F1 VS chrf	.071	
Freq;F1 VS DA	.786*	N/A
Freq;Own VS BLEU	-.071	-.636
Freq;Own VS chrf	-.143	
Freq;Own VS DA	.571	N/A

Table 3: Pairwise Kendall's Tau correlations between our metrics and the standard metrics (DA for direct assessment). The columns are arranged the same way as in Table 2; the statistical significance pointed by asterisk is p < 0.05 (for positive tau values only). For 2022 data, we do not have DA scores, thus it is marked "N/A"; also the rankings by BLEU and chrf are same, thus the corresponding cells in 2022 are merged.

Unfortunately, we cannot compare the 2022 results with human scores yet. For the 2022 automatic scores, the discrepancy between our metric and automated metrics is even bigger, which is represented by the negative τ value. If we analyze the ranking of the systems by the standard metric and of the proposed metrics, we can see that, for 2021, the tentative clustering into three groups (best-average-worst system) roughly coincides with the automated metrics, while for 2022 the general coincidence remains, but there are counterexamples such as Online-W which is best by BLEU

and chrf, and average by our metric. We can interpret the lack of correlation between the automated metrics and our metric the following way: the proposed metrics can give additional information compared to the dominant automated ones; moreover, they tend to correlate with the human document-level judgments, which are, as it has already been mentioned, more sensitive to the inconsistencies in translations on the document level.

5.3 Comparison of 2021 and 2022 Performance

The last notable comparison is the progress of systems that participated in both the 2021 and 2022 competitions; there were six such systems. We subtracted the 2021 scores from the 2022 scores and ranked the differences from the most significant increase to the biggest decrease. We did that both for our metrics and for the standard automatic ones. The first notable difference is that the changes in scores with our metrics are very small compared to BLEU and chrf, they are not bigger than 3% (while the smallest change in BLEU and chrf are 15% and 10%, correspondingly). Based on that, we may hypothesize that our metrics show that the system developers did not aim at increasing the term consistency of the translations. However, to check this hypothesis, we should analyze the architecture of the systems and possibly to compare their performance against the systems intentionally oriented at term preservation, such as Voita et al. (2019a). The detailed comparison of 2021 and 2022 algorithms is shown in Table 4.

6 Limitations and Perspectives

As was stated, we proceed with testing our metrics, both "extensively" (on more data) and "intensively" (by tweaking the inner parameters of the metric itself). Regarding the "extensive" analysis, we firstly should retrieve the automatic metrics obtained for the ELITR agreement corpus and compare them to our findings. Secondly, we should test our method on other language pairs or at least on the opposite English-to-Czech direction.

The second priority covers a more "intensive" analysis of the metric. The method that we suggest is based on several automated (or semi-automated) steps. For each of the steps (keyword extraction, word alignment, manual restriction of the term translations), different approaches and algorithms can be used. So far, we have tested the YAKE Campos et al. (2018) and KeyBERT[4] keyword extractors for the first step. We compared their performance on the legal text outside the main ELITR collection (this means that, for regex-based extractor, we created the templates based on the ELITR connection and applied it to the testing text). Tentative analysis shows that for the considered text, regex term extractor demonstrates the best performance, with 100% precision and 64% recall (7 out of 11 terms). Both YAKE and KeyBERT output an excessive number of false positive results, thus showing a dramatic decrease in precision (best performance – YAKE with 1-token keyword retrieval, 35%). The recall scores for these algorithms decrease as well: the comparable result is performed only by YAKE (54% for 1-token keyword retrieval), while the 2-token length YAKE shows 36% and KeyBERT shows 9%.

This can lead us to the conclusion that the regex term extraction is the best algorithm. However, if we apply these extractors to different texts of a similar domain – audit report (retrieved from another ELITR repository,[5]) we will see that the regex keyword extraction outputs no terms at all. The reason is that the terms in this report are introduced only in parentheses, with no additional explicit hints (such as "hereinafter referred as...") in the legal texts. Both machine learning-based algorithms, in contrast, manage to catch at least some of the necessary terms. This drives us to the conclusion, that for the robustness of the regex-based term extraction, we should take into account different "strategies" of introducing the terms in the document (sometimes – by parentheses, sometimes – by additional phrases). This means that, before evaluating a new collection of the exact text, we still need some human effort to understand the strategy of the term marking there. Another way for a bigger automatization can be using the combination of different keyword extraction algorithms, and choosing the terms through a majority vote or taking the union. Finally, we can look at the problem of the term extraction and alignment from an opposite perspective: if there is no reliable combination of the automated algorithms for these two steps, we can use our metric semi-manually: the steps 2-3 from Section 3 will be completely handed over to human annotators, and their results will be processed automatically

[4] https://maartengr.github.io/KeyBERT/index.html

[5] https://github.com/ELITR/wmt20-elitr-testsuite

	1st; F1	1st; Own	Freq; F1	Freq; Own	1st; F1 rank	1st; Own rank	Freq; F1 rank	Freq; Own rank	BLEU	chrf	BLEU rank	chrf rank
CUNI-Doc Transformer	.0006	.0012	.0006	.0013	3	3	3	3	.1603	.1109	5	4
Online-A	-.0065	-.0219	.0099	.0065	5	5	1	2	.1985	.1371	2	2
Online-B	.0223	.0395	-.0126	-.0097	1	1	5	5	.1749	.1197	3	3
Online-G	-.0005	.0006	-.0012	-.0049	4	4	4	4	.1533	.1031	6	6
Online-W	.0081	.0085	.0051	.0066	2	2	2	1	.2733	.1835	1	1
Online-Y	-.0399	-.0465	-.0288	-.0305	6	6	6	6	.1668	.1078	4	5

Table 4: Comparison of systems' progress from 2021 to 2022. The columns with the names of metrics (or the variants of our metric) denote the result of subtraction of the 2022 scores from 2021 scores. The "rank" columns sort the systems by their progress in the corresponding metric (1 - biggest increase, 6 - lowest increase/biggest decrease).

by steps 4-5. Of course, such implementation will be more time- and effort-consuming, but, firstly, it should still be faster than other manual evaluation approaches such as MQM, secondly, it will give us a model results of term extraction and alignment, against which we will compare the automated algorithms.

The last notable limitation of the proposed approach is rooted in linguistic issues. Although the legal texts are very consistent in using the same term for the same concept, there regularly appear cases of "legitimate" homonymy, where two terms can denote the same concept. This usually occurs when two or more antecedents can be referred to separately or by one term. The example is the following sentence: *X, hereinafter referred to as "Seller", and Y, hereinafter referred to as "Buyer", together also as "contracting parties"....* Such ambiguity (when person X can be both referred as "Seller" and as "contracting parties") may cause the problems even within the correct translation, if in the original the chosen formulation would be "the Seller and the Buyer", and in the target language it would be chosen as "contracting parties". The current metric does not have any capacity to capture this feature of the legal language domain.

7 Conclusion

We have presented the metric for evaluating the terminology consistency of the automatically translated texts. Among its main advantages is its ability to be automatized and its relative simplicity of interpretation. We have tested our metric on the texts from the legal domain in the Czech-to-English translation pair, and we have obtained the results that, according to preliminary estimates, correlate with human document-level judgements and statistically differ from those of the automated metrics such as BLEU or chrF. We are continuing our analysis to understand the scope of our metric's functionality and test it on other language pairs.

We publish our code of the project online at the Github page[6] of the Institute of Formal and Applied Linguistics, Charles University. We will appreciate feedback on the current algorithm, and we are open to discussion and suggestions on its improvement.

8 Acknowledgements

This work was supported by GAČR EXPRO grant LUSyD (20-16819X, RIV: GX20-16819X) and we used services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference

[6] https://github.com/ufal/wmt22-term-based-metric

on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina. 2021. On the evaluation of machine translation for terminology consistency. *CoRR*, abs/2106.11891.

Niyati Bafna, Martin Vastl, and Ondřej Bojar. 2021. Constrained decoding for technical term retention in english-hindi mt. In *Proc. 18th International Conference on Natural Language Processing, ICON 2021, December 16-19, 2021*. NLP Association of India (NLPAI).

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. Yake! collection-independent automatic keyword extractor. In *Advances in Information Retrieval*, pages 806–810, Cham. Springer International Publishing.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Christian Hardmeier. 2012. Discourse in statistical machine translation: A survey and a case study. *Discours*, 11.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Tereza Vojtěchová, Michal Novák, Miloš Klouček, and Ondřej Bojar. 2019. SAO WMT19 test suite: Machine translation of audit reports. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 481–493, Florence, Italy. Association for Computational Linguistics.

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. 2020. WMT20 document-level markable error exploration. In *Proceedings of the Fifth Conference on Machine Translation*, pages 371–380, Online. Association for Computational Linguistics.