PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 108 JUNE 2017 121–132

Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation

Filip Klubička,[a] Antonio Toral,[b] Víctor M. Sánchez-Cartagena[c]

[a] University of Zagreb
[b] University of Groningen
[c] Prompsit Language Engineering

Abstract

We compare three approaches to statistical machine translation (pure phrase-based, factored phrase-based and neural) by performing a fine-grained manual evaluation via error annotation of the systems' outputs. The error types in our annotation are compliant with the multidimensional quality metrics (MQM), and the annotation is performed by two annotators. Inter-annotator agreement is high for such a task, and results show that the best performing system (neural) reduces the errors produced by the worst system (phrase-based) by 54%.

1. Introduction

A paradigm to machine translation (MT) based on deep neural networks and usually referred to as neural MT (NMT) has emerged in the past few years. This has disrupted the MT field as NMT, despite its infancy, has already surpassed the performance of phrase-based MT (PBMT), the mainstream approach to date.

We have witnessed the potential of NMT in terms of overall performance scores, be those automatic (e.g. BLEU) or human (e.g. system rankings); for example, in last year's news translation shared task at WMT.[1] There, out of 9 language directions where NMT systems were submitted, they significantly outperformed PBMT in 8, according to the human evaluation. In the remaining language direction (Russian-to-

[1] http://www.statmt.org/wmt16/translation-task.html

Corresponding author: fklubick@ffzg.hr
Cite as: Filip Klubička, Antonio Toral, Víctor M. Sánchez-Cartagena. Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 121–132. doi: 10.1515/pralin-2017-0014.

English), the best PBMT submission was ranked higher than the best NMT system, but the difference was found not to be significant.

Given the impressive overall performance of NMT, some researchers have attempted in the past year to analyse the potential of NMT in a more detailed manner. The motivation comes from the fact that while overall scores give an indication of the general performance of a system, they do not provide any additional information. Hence, in order to delve further and try to shed light on the strengths and weaknesses of this new paradigm to MT, two recent papers have looked at conducting multifaceted evaluations.

- Bentivogli et al. (2016) conducted a detailed analysis for the English-to-German language direction where they compared state-of-the-art PBMT and NMT systems on transcribed speeches. They found out that NMT (i) decreases post-editing effort, (ii) degrades faster than PBMT with sentence length and (iii) improves notably on reordering and inflection.
- Toral and Sánchez-Cartagena (2017) carried out a series of analyses and evaluations for NMT and PBMT systems on the domain of news for 9 language pairs. They corroborated the findings of Bentivogli et al. (2016) with respect to NMT outstanding performance on reordering and inflection and its degradation with sentence length. They also contributed additional findings: NMT systems (i) exhibit higher inter-system variability, (ii) lead to more fluent outputs and (iii) perform more reordering than PBMT but less than hierarchical PBMT.

A limitation of these analyses lies in the fact that all of them were performed automatically. E.g. reordering and inflection errors were detected based on automatic evaluation metrics. Hence, one could argue that their outcomes are somewhat affected as automatic tools are, of course, never perfect.

In this paper we conduct a detailed human analysis of the outputs produced by NMT and PBMT systems. Namely, we annotate manually the errors found according to a detailed error taxonomy, that is compliant with the hierarchical listing of issue types defined as part of the Multidimensional Quality Metrics (MQM) (Lommel et al., 2014a). Specifically, we carry out this analysis for the news domain in the English-to-Croatian language direction. First, we define an error taxonomy that is relevant to the problematic linguistic phenomena of this language pair. Subsequently, we annotate the errors produced by 3 state-of-the-art translation systems that belong to the following paradigms: PBMT, factored PBMT and NMT. Finally, we analyse the annotations.

The main contributions of this paper can then be summarised as follows:

1. We conduct, to the best of our knowledge, the first human fine-grained error analysis of NMT in the literature.
2. We analyse NMT in comparison not only to pure PBMT and hierarchical PBMT, as in previous works, but also with respect to factored models.
3. We develop an MQM-compliant error taxonomy for Slavic languages.
4. We develop a novel approach to statistically analyzing and interpreting the results of MQM error annotation.

The rest of the paper is organised as follows. Section 2 describes the MT systems and the datasets used in our experiments. Section 3 covers the analysis, including the definition of the error taxonomy, the annotation setup and guidelines and finally the results obtained and their discussion. Finally, Section 4 outlines the conclusions and lines of future work.

2. MT Systems

This section describes the MT systems and the datasets used in our experiments. We built PBMT, factored PBMT and NMT systems.

The 3 systems were trained on the same parallel data. We considered a set of publicly available English–Croatian parallel corpora, comprising the DGT Translation Memory [2], HrEnWaC[3], JRC Acquis [4], OpenSubtitles 2013, SETimes and Ted talks. We concatenated all these corpora and performed cross-entropy based data selection (Moore and Lewis, 2010) using the development set. Once the data is ranked we keep the highest ranked 25% sentence pairs (4,786,516).

PBMT systems used also monolingual data for language modelling. To this end we used the concatenation of the hrWaC corpus (Ljubešić and Klubička, 2014) and the target side of the aforementioned parallel corpora.

As development set we used the first 1,000 sentences of the English test set used at the WMT12 news translation task[5], translated by a professional translator into Croatian. Similarly, our test set is made of the first 1,000 sentences of the English test set of the WMT13 translation task[6], again manually translated into Croatian.

The PBMT system was built with Moses v3.0[7]. In addition to the default models we also used hierarchical reordering (Galley and Manning, 2008), an operation sequence model (Durrani et al., 2011) and a bilingual neural language model (Devlin et al., 2014).

The factored PBMT system maps one factor in the source language (surface form) to two factors in the target (surface form and morphosyntactic description). This system is described in detail by Sánchez-Cartagena et al. (2016).

The NMT system is based on the sequence-to-sequence architecture with attention and we applied sub-word segmentation with byte pair encoding (Sennrich et al., 2015) jointly on the source and target languages. We performed 85 000 join operations. Training was run for 10 days and a model was saved every 4.5 hours. We decoded

[2] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

[3] https://www.clarin.si/repository/xmlui/handle/11356/1058

[4] http://tinyurl.com/CroatianAcquis

[5] http://www.statmt.org/wmt12/translation-task.html

[6] http://www.statmt.org/wmt13/translation-task.html

[7] https://github.com/moses-smt/mosesdecoder/tree/RELEASE-3.0

the test set using an ensemble of 4 models. These were the 4 models with the highest BLEU scores on the development set.

2.1. Evaluation

We report the scores obtained in terms of the BLEU and TER automatic evaluation metrics for the 3 systems described in the previous section. Table 1 shows the results.

As the table shows, the use of factored models leads to a substantial improvement upon pure PBMT (6% relative in terms of BLEU). NMT, on its turn, allows us to obtain a further notable improvement; 14% relative in terms of BLEU compared to the factored PBMT system and 21% compared to the initial PBMT system.

System	BLEU	TER
PBMT	0.2544	0.6081
Factored PBMT	0.2700	0.5963
NMT	0.3085	0.5552

Table 1. Automatic evaluation (BLEU and TER scores) of the 3 MT systems

3. Error analysis

In this section we report on the motivation for conducting the manual error analysis, describe the framework and overall annotation process, and present the results.

The fact that Croatian is rich in inflection, has rather free word order and other similar phenomena that English does not, gives rise to specific translation issues. For example, grammatical categories that do not exist in English, like gender and case, may be particularly hard to generate reliably in a Croatian translation. We built our factored PBMT system aiming to directly address such issues. Similarly motivated, we wished to see how an NMT system would grapple with the same issues.

Indeed, as shown in Section 2, automatic evaluation shows significant improvement for both systems, compared to the pure PBMT system. However, as is the nature of automatic metrics, the automatic scoring methods do not indicate whether any of the linguistic problems mentioned earlier have been addressed by the systems. The question of whether the linguistic quality, or rather, grammaticality of the output is improved has not been answered by automatic evaluation. Are cases and gender handled better? Is there better agreement? Is the fluency of the translation higher?

In order to provide answers to these research questions, we decide to thoroughly compare these systems by systematically analyzing their outputs via manual error analysis. In this way we can obtain a more complete picture of what is happening in the translation, which can provide pointers on where to act to obtain further improvements in the future.

3.1. Multidimensional Quality Metrics and the Slavic tagset

After looking into different ways of performing the task of manual evaluation via error analysis, we decided to make use of the MQM framework, developed in the QT-Launchpad project[8]. This is a framework for describing and defining custom translation quality metrics. It provides a flexible vocabulary of quality issue types and a mechanism for applying them to generate quality scores. It does not impose a single metric for all uses, but rather provides a comprehensive catalog of quality issue types, with standardized names and definitions, that can be used to describe particular metrics for specific tasks.

The main reason we chose the MQM framework was the flexibility of the issue types and their granularity — it gave us a reliable methodology for quality assessment, that still allowed us to pick and choose which error tags we wish to use.

The MQM guidelines propose a great variety of tags on several annotation layers[9]. However, the full tagset is too comprehensive to be viable for any annotation task, so the process begins with choosing the tags to use in accordance to our research questions. Initially we started off with the core tagset, a default set of evaluation metrics (i.e. error categories) proposed by the MQM guidelines, as seen in Figure 1.

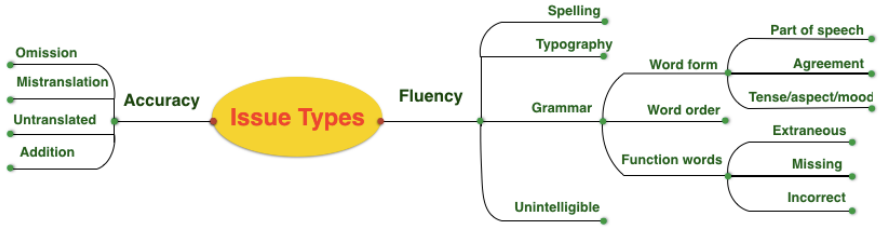

Figure 1. The core error categories proposed by the MQM guidelines

However, given the morphological complexity of Croatian and the level at which we made interventions in the system, we found that these core categories were not detailed enough, or rather, did not allow for an analysis of the specific phenomena we were interested in. Some categories that were of interest to us, like specific *Agreement* types, were not present in the tagset, while some errors, like *Typography*, were irrelevant to us. So we created our own set of tags by modifying the core set, rearranging the hierarchy, adding new tags and removing those that are of little relevance. We call this new tagset the Slavic tagset, as its expansion allows for the identification of

[8] http://www.qt21.eu/mqm-definition/definition-2015-06-16.html

[9] http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html

grammatical errors which are commonly shared by Slavic languages. This tagset is outlined in Figure 2.

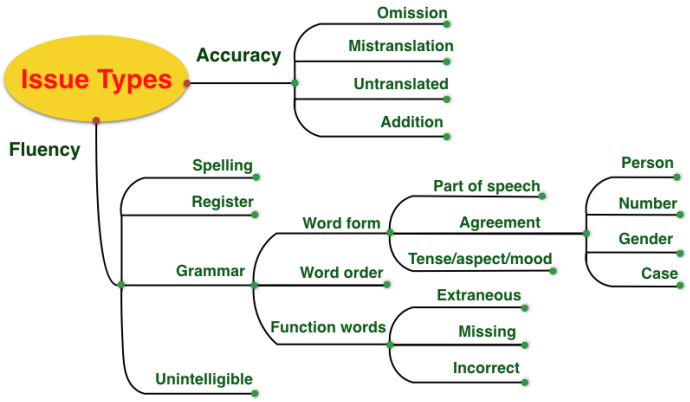

Figure 2. The Slavic tagset, a modified version of the MQM core tagset

3.2. Annotation setup

In order to carry out the annotations we used `translate5`[10], a web-based tool that implements annotations of MT outputs using hierarchical taxonomies, as is the case of MQM.

We had two annotators at our disposal, who both had prior experience with MQM as well as the same background - an MA in English linguistics and information science. They were thoroughly familiarized with the official annotation guidelines and the decision process[11] prior to annotation.

The annotators annotated 100 random sentences from the test set introduced in Section 2. These sentences were translated by all three MT systems, and the annotators were presented with the source text, a reference translation and the unannotated system outputs at the same time. All three translations were then annotated by both our annotators (i.e. each system translated the same 100 sentences, each annotator annotated the 300 translated sentences, making a total of 600 annotated sentences). Once the sentences were annotated, the annotation data was extracted, we calculated inter-annotator agreement and analyzed the output to see what the number of error tags can tell us about the performance of each system.

[10] `http://www.translate5.net/`

[11] `http://www.qt21.eu/downloads/annotatorsGuidelines-2014-06-11.pdf`

3.3. Inter-Annotator Agreement

Though carefully thought out and developed, the MQM metrics, and manual MT evaluation in general, are notorious for resulting in low inter-annotator agreement scores. This is attested by the body of work that has addressed this issue, most notably Lommel et al. (2014b), who worked specifically on MQM, and (Callison-Burch et al., 2007), who investigated several tasks. This is why it is important that we check how well our annotators agree on the task at hand, and whether this is consistent with other work done with MQM so far.

Once the data was annotated, agreement was observed at the sentence level, and inter-annotator agreement was calculated using the Cohen's Kappa (κ) metric (Cohen, 1960). Agreement was calculated on the annotations of every system separately, as well as on a concatenation of annotations, in order to both see whether there are differences in agreement across systems, as well as to gain insight into the overall agreement between annotators. Additionally, Coehn's κ was also calculated for every error type separately. Detailed results can be found in Table 2.

Generally, one can see that our annotators agree best on evaluations of the PBMT system, less so on evaluations of the Factored SMT system, and least in evaluations of the NMT system. Overall agreement scores are relatively low - the average total κ is approximately 0.51. Furthermore, the κ scores are relatively consistent across all error types, mostly ranging between 0.35 and 0.55. According to Cohen, such scores constitute moderate agreement. However, as already stated, this is to be expected, given the complexity of the problem and annotation schema. In fact, this is a notably higher score than what has been reported in similar work, e.g. Lommel et al. (2014b), who achieve κ scores ranging between 0.25 and 0.34. However, this comparison should be taken with a grain of salt, as our calculations are just an approximation compared to Lommel et al.'s, given that in our setup we looked only at sentence level agreement, while they calculated agreement on the token level.

3.4. Results of annotation

Directly extracting raw annotation data from the `translate5` system provides a sum of error tags annotated for each error type by each annotator and system. The total values are presented in Table 3.

Looking at the aggregate data alone, one can easily detect that both annotators have judged that the PBMT system contains the most errors, and that the NMT system contains the smallest number of errors. This trend is consistent across most fine-grained error categories as well.

However, even though simply counting the errors can provide insight into which system performs better, we thought that this approach does not adequately represent our findings, as it does not allow a proper quantification of the quality of the outputs. Certainly, based on data from Table 3 we can claim, for example, that the NMT system

Error type	PBMT	Factored	NMT	Concatenated
Accuracy				
Mistranslation	0.51	0.48	**0.58**	0.53
Omission	0.34	0.39	0.37	0.37
Addition	0.5	0.54	0.33	0.47
Untranslated	**0.86**	**0.86**	-0.02	**0.72**
Fluency				
Unintelligible	0.39	0.32	0	0.35
Register	0.37	0.2	0.22	0.27
Word order	0.56	0.33	0.21	0.4
Function words				
Extraneous	0.56	0.32	0.49	0.46
Incorrect	0.37	0.18	0.34	0.29
Missing	0	0.49	0	0.33
Tense...	0.44	0.36	0.15	0.38
Agreement	0.24	0.41	0	0.33
Number	0.53	0.55	0.52	0.54
Gender	0.46	0.59	0.48	0.53
Case	0.53	0.49	0.52	0.56
All errors	**0.56**	**0.49**	**0.44**	**0.51**

Table 2. Inter-annotator agreement (Cohen's κ values) for the MQM evaluation task. The highest score for any individual system and the concatenation, as well as the overall score, are shown in bold.

	Annotator 1			Annotator 2		
System	**PBMT**	**Factored**	**NMT**	**PBMT**	**Factored**	**NMT**
Total errors	317	276	178	264	199	132

Table 3. Total errors per system per annotator

produces less errors in general, or less errors of a specific type, but given that the outputs are different, as is the number of tokens in each translation, we decided to normalize the data.

To the best of our knowledge there is no related work on how to approach this, as previous work simply counts the number of MQM tags and stops there. After some consideration, we decided to normalize at the token level. I.e. instead of counting just error tags produced by each annotator, we count the tokens that these errors are assigned to – tokens that do and tokens that do not have an error annotation. Once

these numbers are divided by the total number of tokens in the system's output, they provide a concrete idea of the ratio of tokens with and without errors.

The results of such analysis again show that the PBMT system has the largest error ratio, while the NMT system has the smallest one. This is further backed up by a pairwise chi-squared (χ^2) statistical significance test; we calculate statistical significance from 2x2 contingency tables for every system pair (PBMTxFactored, PBMTxNMT and FactoredxNMT). The results show that the differences in the total number of tokens with errors are statistically significant for all three system pairs, with the p value being lower than 0.0001 in each case.

Furthermore, we also wanted to see which error types are the ones making a significant impact on this result. So we repeated these same measurements, but instead of performing them on all error types combined, they were performed separately for each specific error category. The combined results of the calculations and transformations are presented in Table 4.

We can derive several findings from this table. Firstly, when looking simply at the grand total of tokens with and without errors, the difference between the systems is statistically significant by a wide margin. When looking at PBMT and factored PBMT, the factored system has significantly less errors than the pure PBMT system. The overall error rate is in this case reduced by 20%. A separate analysis of specific error types that contribute to this score reveals that only some of the error categories are significantly different between the two systems. In the table, those categories are filled in with green. One can see that, when it comes to agreement, the only agreement type that produces significantly less errors is agreement in case.

However, taking a look at NMT shows that, not only does it result in a 42% overall error reduction compared to the factored system, and 54% with respect to pure PBMT, but it produces even less agreement errors – overall, as well as at the level of number, gender and case – while not using any kind of linguistic information at all. This might in part be due to the use of sub-word segmentation, as inflections in Croatian are relatively regular. In addition to improving in the Agreement category, NMT also produces significantly less errors in many more categories than the factored model does. Interestingly, it produces more Omission errors than either of the other two systems. It seems that it tends to sacrifice completeness of translation in order to increase overall fluency. Indeed, extrapolating from the data in Table 4, shows that, though differences are very small, NMT does have the lowest token per sentence ratio (PBMT 18.99, Factored PBMT 18.89, NMT 18.36).

4. Conclusion

The fine-grained manual evaluation performed for the purpose of this research has provided answers to several questions, one of which was the main drive behind our developing the factored system: is there a way to handle better agreement when

	PBMT		Factored		NMT	
Error type	**No error**	**Error**	**No error**	**Error**	**No error**	**Error**
Accuracy	3467	369	3525	*291	3402	266
Mistranslation	3547	289	3586	*230	3471	197
Omission	3801	35	3793	23	3619	*49
Addition	3814	22	3797	19	3655	13
Untranslated	3813	23	3797	19	3662	*6
Fluency	3195	641	3298	*518	3465	**188
Unintelligible	3790	46	3769	47	3668	**0
Register	3810	26	3794	22	3646	22
Spelling	3833	3	3812	4	3659	9
Grammar	3270	566	3371	**445	3497	**156
Word order	3752	84	3752	64	3646	**22
Function words	3801	35	3780	36	3650	*18
Extraneous	3829	7	3810	6	3664	4
Incorrect	3810	26	3790	26	3655	*13
Missing	3834	2	3812	4	3667	1
Word form	3389	447	3471	*345	3538	**102
Part of speech	3822	14	3800	16	3663	*5
Tense...	3775	61	3765	51	3648	*20
Agreement	3466	370	3540	*276	3566	**102
Number	3778	58	3772	44	3646	*22
Gender	3788	48	3756	60	3644	*24
Case	3614	222	3694	*122	3622	**46
Person	3836	0	3816	0	3664	4
Total errors	2826	1010	3007	**809	3199	**469

*Table 4. Processed annotation data from both annotators concatenated: each system's total number of tokens with and without errors. Statistical significance for a system, when compared to the system on its left, is marked with * where* p*-value is <0.05 and ** where* p*-value is <0.0001. Cells with a green background indicate that the system has less errors than the one on its left, while those in red indicate that it has more.*

translating to Croatian? We can now confidently claim that factored models result in significantly less agreement errors overall compared to pure PBMT.

We can also confidently claim that NMT handles all types of agreement better than both pure PBMT and factored PBMT, which corroborates the findings of other researchers' NMT evaluations. Our system produces sentences with far less errors, and a language that is more fluent and more grammatical, which should be of help when it comes to the task of post-editing.

Furthermore, the error taxonomy that was developed for this research, while only used for the English-to-Croatian language direction, should be applicable for the analysis of errors for any translation direction towards a Slavic language, as it takes into account grammatical properties specific to these languages.

Among other possible lines of future work, including the application of our methodology to another language pair (e.g. English-Czech), performing more controlled IAA analysis or IAA adjudication, as well as comparing to an NMT model without sub-word segmentation, another one is adapting the tagset further. In its current version, it has proved to be informative when comparing PBMT to factored PBMT. However, NMT has shown itself to produce language that is so fluent that the fine-grained hierarchy in the *Fluency* branch is of little use. Meanwhile, the most common error type in the NMT output is *Mistranslation*, which, according to the MQM guidelines, covers both lexical selection and, less intuitively, translation of grammatical properties (e.g. if 'cats[pl.]' is translated as 'mačka[sg.]', this is to be tagged as *Mistranslation*, in spite of correct lexical choice). This makes it quite a vague category, so if one would wish to perform an even more nuanced linguistic error analysis for NMT, adding additional layers to the *Accuracy* branch would seem a promising direction to follow.

Acknowledgements

We would like to extend our thanks to Maja Popović, who provided valuable advice on how to approach the annotation and evaluation, and Denis Kranjčić, who participated in the annotation task. The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran) and the Swiss National Science Foundation grant 74Z0_160501 (ReLDI).

Bibliography

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267. Association for Computational Linguistics, 2016.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158. Association for Computational Linguistics, 2007.

Cohen, Jacob. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Devlin, Jacob, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *Proceedings of ACL*, pages 1370–1380, 2014.

Durrani, Nadir, Helmut Schmid, and Alexander Fraser. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1045–1054. Association for Computational Linguistics, 2011.

Galley, Michel and Christopher D Manning. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856. Association for Computational Linguistics, 2008.

Ljubešić, Nikola and Filip Klubička. {bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

Lommel, Arle Richard, Aljoscha Burchardt, and Hans Uszkoreit. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica: tecnologies de la traducció*, 0(12):455–463, 12 2014a.

Lommel, Arle Richard, Maja Popovic, and Aljoscha Burchardt. Assessing inter-annotator agreement for translation error annotation. In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*, 2014b.

Moore, Robert C. and William Lewis. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Sánchez-Cartagena, Victor M., Nikola Ljubešić, and Filip Klubička. Dealing with Data Sparseness in SMT with Factured Models and Morphological Expansion: a Case Study on Croatian. *Baltic Journal of Modern Computing*, 4(2):354–360, 2016.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Toral, Antonio and Víctor M. Sánchez-Cartagena. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. *CoRR*, abs/1701.02901, 2017.

Address for correspondence:
Filip Klubička
fklubick@ffzg.hr
Faculty of Humanities and Social Sciences
3 Ivan Lučić Street, Zagreb, PA 10000, Republic of Croatia