Human Evaluation of Machine Translations by Semi-Professionals: Lessons Learnt

Špela Vintar*, Andraž Repar†
* Department of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana
spela.vintar@ff.uni-lj.si

†Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
andraz.repar@ijs.si

Abstract

We report on two experiments in human evaluation of machine translations, one using the Fluency/Adequacy scoring and the other using error annotation combined with post-editing. In both cases the evaluators were students of translation at the Master's level, who received instructions on how to perform the evaluation but had previously had little or no experience with the evaluation of translation quality. The human evaluation was performed in the context of development and testing different MT models within the Development of Slovene in a Digital Environment (DSDE) project. Our results show that Fluency/Adequacy scoring is more efficient and reliable than error annotation, and a comparison of both methods shows low correlation.

1. Introduction

The design and evolution of a new machine translation system is invariably linked with regular quality assessments, using both automatic methods commonly known as metrics and human evaluations of the MT system's outputs. The context of this experiment is the development of a neural MT system for the English-Slovene language pair within the DSDE project, which involved work packages dedicated to data collection, implementation and testing of different NMT architectures and MT evaluation.

Throughout the project, different versions of the DSDE NMT system were regularly automatically evaluated using the BLEU metric, while later versions were also evaluated with a comprehensive set of scores available on the SloBench 1.0 evaluation platform. In parallel to the automatic ones we performed a set of human evaluations with several aims in mind: To validate the automatic scores with manual assessments, to gain insight into the performance of the system under development, but also to compare two human evaluation scenarios in terms of efficiency and reliability.

The manual evaluations of the DSDE MT engine were performed by students of MA Translation at the Department of Translation Studies, Faculty of Arts, University of Ljubljana. We refer to advanced students of translation as semi-professionals because of their high proficiency in both languages and their understanding of translation as a complex cognitive activity with many alternative solutions for each source text. On the other hand, their experience with translating is for the most part limited to the study environment, and they have received little or no formal training in post-editing or translation assessment.

Manual evaluation was performed using two common evaluation frameworks: the Adequacy/Fluency score and the MQM-DQF error annotation combined with post-editing.

The paper first presents the rationale for selecting the methodologies by referring to related work, then describes the MT system and its development within the DSDE project. We then present the evaluation setups and provide summaries of the results. In addition to quantitative results, for the error annotation and post-editing task we also give a brief summary of the most frequent observations. We conclude by discussing the findings from the perspective of translation quality assessment in MT development.

2. Related work

Evaluation of MT is a crucial part of development and improvement of MT systems, and it is traditionally divided into automatic evaluation using metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), and human or manual evaluation. Automatic evaluation is usually performed by comparing the candidate machine translated text to a reference translation produced by a human professional, whereby the comparison can be rather superficial and word-based such as with BLEU, or more linguistically informed such as with METEOR. The obvious advantage of automatic metrics is that they can be performed on the fly requiring no human effort, but the rate of correlation with human judgements remains a constant concern. Particularly since the emergence of NMT, some authors show that the reliability of metrics as indicators of translation quality may be faltering (Shterionov et al., 2018), or that metrics alone cannot adequately reflect the variety of linguistic issues which may affect quality. Manual evaluation therefore remains an integral part of MT quality evaluation and is annually included into the WMT shared task (Bojar et al., 2016).

Over time, many methods of human MT evaluation have evolved. The Adequacy/Fluency scoring was first adopted by the ARPA MT research program (White et al., 1994) as a standard methodology of scoring translation segments on a discreet 5- or 7-level scale. The adequacy evaluation is performed either by professional translators who are presented with the original and the machine translated segment and make judgments about the degree to which the information from the original can be found in MT output, or by monolingual speakers who are presented with the MT and a reference translation. For the fluency evaluation, no reference translation nor original is provided and the evaluators determine whether the translation

"reads" like good language, sounds natural and adheres to grammatical conventions of the language.

Other manual evaluation methods include task-based evaluation (Doyon et al., 1999), post-editing with targeted human annotation, also known as HTER (Snover et al., 2006), and error analysis using various error typologies. The most comprehensive translation error typology to date is the Multidimensional Quality Metrics (MQM) developed in the QT-Launchpad[1] project. The MQM guidelines provide a fine-grained hierarchy of quality issues and a mechanism for applying them to different evaluation scenarios, however the entire tagset is too complex to be used in a concrete evaluation task. Originating from the needs of the language industry, the TAUS Dynamic Quality Framework (DQF) proposed an error typology which has been harmonized with the MQM model in 2015 and is today integrated into most commercial translation tools (Marheinecke, 2016).

The annotation of translation errors can be a part of Linguistic Quality Assurance (LQA) in professional translation environments, in order to monitor quality on the corporate, project or individual levels. However, for the task of manual MT evaluation MQM and related methods are notoriously poor in inter-annotator agreement scores (Lommel et al., 2014). Some authors believe that pre-annotation training can significantly reduce disagreements, but the task apparently remains highly subjective.

Despite the labour intensity and low inter-annotator agreement, error annotation is still frequently employed in human MT evaluation because of the significance and depth of insight into translation issues it may provide. As Klubička et al. (2017) point out, Slavic languages are rich in inflection, case and gender agreement, and they have rather free word order compared to English. The motivation for using error analysis in MT evaluation is to see – in the process of developing and improving a new MT system – whether the particular grammatical issues occurring with Slavic languages are adequately addressed, resulting in overall quality improvement.

In line with related works we opted for two of the most commonly used manual evaluation methods, the Fluency/Adequacy score and the TAUS DQF-MQM metrics which has been further adapted for the DSDE project.

3. The DSDE MT system

The main goal of the machine translation work package in RSDO is to improve on the state-of-the-art model for the Slovene/English and English/Slovene language pairs developed within the TraMOOC project (Sennrich et al., 2017). To this end, various neural machine translation frameworks were evaluated, such as *MarianNMT* (Junczys-Dowmunt et al., 2018), *fairseq* (Ott et al., 2019) and *NeMo* (Kuchaiev et al., 2019). The same dataset consisting of publicly available parallel data as well as data collected within the DSDE project[2] was used to train the models on the selected frameworks.

4. Evaluation setup

Both types of manual evaulation were performed by students of MA Translation at the Department of Translation, University of Ljubljana. The translation environment of choice was memoQ, a tool which allows the project manager to select or define an LQA scheme with the fluency/adequacy scoring or the error categories respectively. The annotator performs the evaluation, error annotation and post-editing in a typical two-column setting with the segmented original on the left hand side and the machine translated segments already inserted into the target text on the right hand side via pre-translation. Annotators receive an outbound memoQ package which ensures that the source text, the raw MT and the evaluation/error annotation scheme are available and activated with no further setup, and the evaluated, post-edited and annotated texts may be returned to the project manager (in our case the experiment designer) as inbound return packages.

Five different source texts were used from the domains of chemistry, karst, economy, law and general news. The texts were of comparable length (~500 words) and consisted either of the entire text or a meaningful portion thereof. With the exception of the general news text dealing with US elections, all domain-specific texts were highly specialized with complex syntax and many terminological expressions.

For the fluency/adequacy scoring, both language pairs were evaluated by a group of five students over a period of several months. Each document was evaluated by two students. Once a new model was available, MemoQ packages were sent to the students who performed the evaluation in their home environment. Note that for the adequacy/fluency evaluation, no postediting took place – the students only had to score each translated segment on a scale of 0 to 3 (see Table 1).

	Adequacy	Fluency
0	None	Incomprehensible
1	Little	Disfluent
2	Much	Good
3	All	Flawless

Table 1: Adequacy/Fluency scoring.

For the error annotation, only the English-Slovene language pair was evaluated, with English as original and Slovene as target. Fifteen students participated, so that post-editing and error analysis were in the end performed by three students for each text. The experiment took place during a regular face-to-face seminar session in the presence of the lecturer. Students were using standard PCs and with memoQ 9.5 running Translator Pro licenses.

Students were requested to perform full post-editing of the machine translated text, and at the same time annotate each error using the preloaded TAUS/DSDE error typology. The latter proved somewhat wearisome, since the annotation of each single error involves opening a separate dialog box, selecting the category and resuming work, whereby the typical commands used during "normal" translation must be avoided (e.g. Control + Enter to confirm the segment). This invariably slows down the post-editing process and presumably affects the natural cognitive flow during post-editing.

[1] https://www.qt21.eu/launchpad/index.html

[2] Data collected within the DSDE project will be made available under a CC-BY-SA 4.0 license at the end of the project.

5. Results

5.1. Fluency/Adequacy scoring

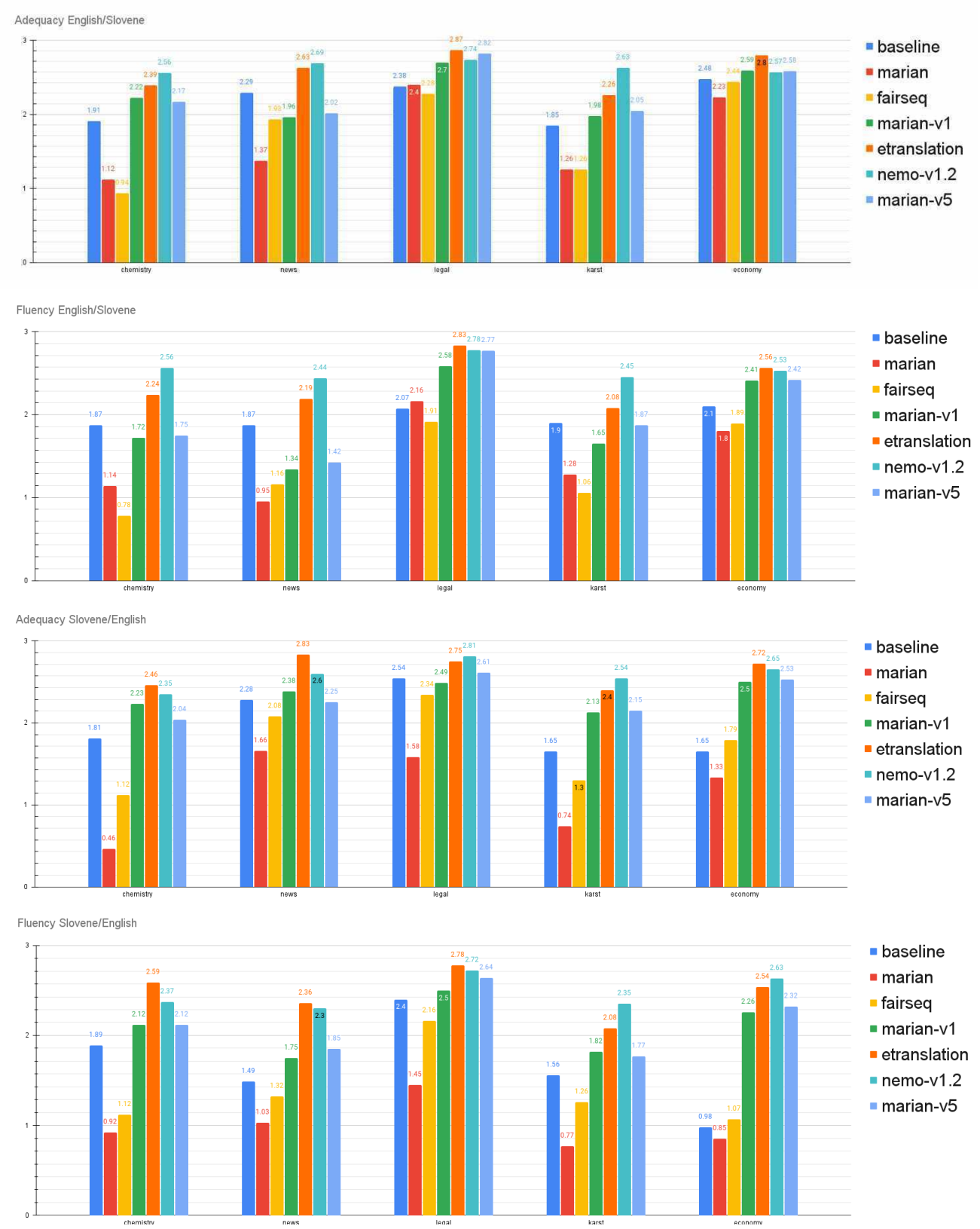

Figure 1. Adequacy and Fluency scores across five domains and two language pairs.

In addition to the baseline model, five models were evaluated using the Adequacy/Fluency methodology (three versions trained using the marianNMT framework, one using the fairseq framework and one using the NeMo framework. We also performed one round of evaluation of the *eTranslation* system developed by the European Commission.

The initial models (*marian* and *fairseq*) performed badly and did not exceed the scores of the baseline model in the DSDE project, but additional iterations performed

better. The overall best performance was exhibited by the NeMo model with best or close to best scores in all five domains. The latest version of the Marian model (*marian-v5*) also performed well in some domains (e.g. Legal) less well in others. When comparing the DSDE models with eTranslation, we can observe that the NeMo model offers competitive performance across all five domains (with the possible exception of the News domain for the Slovene/English language pair).

5.2. Error annotation with post-editing

The error annotation with post-editing was performed in order to gain insight into the translation issues most affecting MT quality, but also to assess the efficiency and reliability of this methodology when used with semi-professional translators. The evaluation took place in November 2021 using the output of the marian-v5 model.

.

Category	Subcategory	Severity 1 - Critical	Severity 2 - Major	Severity 3 - Minor
Accuracy	Category total	**56**	**68**	**37**
	Addition	1	2	3
	Mistranslation	50	63	30
	Omission	5	3	4
Language	Category total	**3**	**26**	**57**
	Grammar	3	18	37
	Spelling	0	8	20
Style	Category total	**13**	**18**	**80**
	Awkward	6	15	55
	Inconsistent	7	3	25
Terminology	Category total	**4**	**16**	**14**
Total		76	128	188

Table 2: Total errors by category.

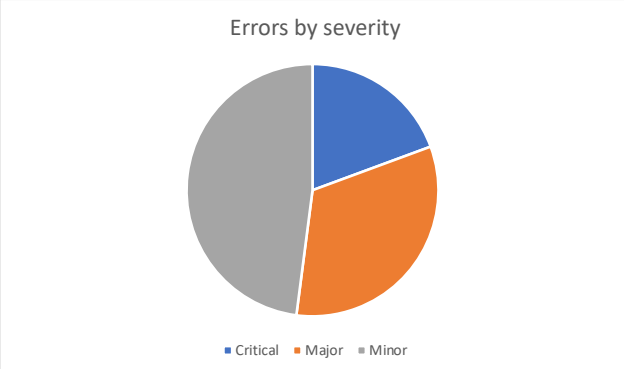

Figure 2: Errors by severity.

As shown in Table 2, the highest number of errors were marked in the Accuracy category, followed by Style, Language and Terminology. Given that four out of five texts were specialized, the low count of terminology errors is perhaps surprising but can be attributed to the fact that annotators frequently choose the Accuracy->Mistranslation category for errors related to specialized lexis. Minor errors are the most frequently selected severity level, with a majority of stylistic errors. Accuracy is also the source of the most critical errors which, in the opinion of annotators, completely change the meaning of the text.

5.2.1. Errors by text

On average, students would annotate ~30 errors per text, or 1.2 errors per segment. The differences in the number of errors between texts are small, with a maximum of 102 errors for the legal text (the sum for all three annotators) and a minimum of 90 for the text on karst.

Category	Subcategory	Chemistry	Economy	Karst	Legal	News
Accuracy	Category total	40	39	58	19	49
	Addition	10	1	0	0	1
	Mistranslation	30	36	56	13	44
	Omission	0	2	2	6	4
Language	Category total	30	14	16	26	18
	Grammar	19	13	15	11	14
	Spelling	11	1	1	15	4
Style	Category total	19	36	13	45	25
	Awkward	19	27	13	22	22
	Inconsistent	0	9	0	23	3
Terminology	Category total	6	5	3	12	9
Total		95	94	90	102	101

Table 3: Errors by text.

Chemistry: There is considerable variation in the number of errors marked by each annotator: 40 / 26 / 29. In all 3 annotations, the most frequent error types are Accuracy and Language, followed by Style and Terminology. Only one annotator found 2 critical errors, the majority of errors were markes as minor.

Economy: The number of errors marked by each annotator varies: 29 / 30 / 35. Similar to other texts, the highest number of errors were attributed to Accuracy->Mistranslation, followed by Style and Language, and only 5 terminology errors.

Karst: The three annotators diverged in the numbers of errors marked: 21 / 31 / 38. Contrary to other texts, here the majority of errors were found to be major or even critical, with only 22 errors categorized as minor. Given that the text was highly specialized, it is again surprising that the Terminology category was not selected more often.

Legal: For the legal text, variation and non-agreement between annotators is at its highest: they marked 21 / 54 / 27 errors each, and even more interesting is the distribution of errors amongst severity levels. For the most prolific annotator, only 4 errors were found to be critical, but for the annotator who spotted 21 errors, 11 were categorized as critical. The third annotator on the other hand found no critical errors.

News: The numbers of errors marked by each annotator were 28 / 33 / 40 respectively, with 12 / 10 / 6 critical errors. Despite the fact that this text was the least specialized of the five, annotators marked 9 errors as terminological, and the overall majority of errors were those pertaining to accuracy (49).

5.2.2. Analysing students' edits

Some texts were highly specialized and rich in terminology, the students however often perceive errors as minor and categorize terminology errors under Accuracy. In the Karst text for example, the original contains the term "precipitation" which is translated as "padavine". None of the annotators identified this as a critical error: in geology, precipitation is not a weather phenomenon but a type of sedimentation process, and the correct translation would read "precipitacija" or "usedanje". The word "test" in the original is most likely a typo and remains untranslated, while the translation of "algal crusts" into "drogovi" is another critical error.

> *In nature, many types of CaCO3 precipitation are linked to living organisms: test, shells, skeletons, stromatolites, algal crusts, etc.*

> *V naravi so številne **padavine** CaCO3 povezane z živimi organizmi: test, oklepi, skeleti, stromatoliti, drogovi itd.*

The students' edits are sometimes unnecessary or even wrong, as in the case of the correctly translated word "adduction" -> "addukcija" corrected into "adukcija" in one case, and in another into "uporaba".

Inconsistent translations are another common issue in machine translation. Thus, in the Economy text, "expenditure" is translated as "stroški", "izdatki", "poraba"; "plant" as "rastlina" and "naprava". A trained and alert post-editor would spot such inconsistencies and make sure they are consolidated in the final version, the students however focus on single segments and overlook such unwanted variation.

Easier to spot are untranslated words, such as "speleothem" in both the Karst original and the Slovene MT. All three annotators spotted the error and opted for "kapnik" in their edits, but the correction is inadequate because "kapnik" is a hyponym of "speleothem" and a better translation would be "speleotem" or "siga". Two annotators marked the error as Critical and one as Major.

It seems that students of translation are much more sensitive to grammatical errors than terminological ones, as the example below containing the correct phrase but in the wrong case was marked as a Major error by all three annotators.

> *Zaradi velike moči odpornosti proti svetlobi in trajnosti derivatov benzimidazolov se pogosto uporabljajo za proizvodnjo **akvarele in elektrofotografskih razvijalnih toner**.*

> *Zaradi velike moči odpornosti proti svetlobi in trajnosti derivatov benzimidazolov se pogosto uporabljajo za proizvodnjo **akvarelnih in elektrofotografskih razvijalnih tonerjev**.*

In many cases the annotators agree on the error itself or the portion of text which should be corrected, but categorize the error differently. A major error was unanimously marked by all three annotators in the Economy text, where the original "To repress these troubles" was machine translated to "Za ponoven tisk te težave". Corrections ranged from "spoprijemanje s težavo", "zmanjšanje teh težav" to "blaženje teh težav", but the error was categorized as Accuracy->Addition, Accuracy ->Mistranslation and Terminology respectively.

Disagreement in categories was frequent also in the non-specialized text, a news article reporting Trump's attempts to postpone elections. The MT version contains a fluent but inaccurate rendering of "November's presidential elections to be postponed", where the MT engine proposed "je predlagal predsedniške volitve v novembru". This is certainly a critical accuracy error, which should be categorized as omission since the postponement was missing in the target. Indeed all three annotators identify the error as critical, but one categorized as mistranslation and the other two as omission. Another severe mistranslation occurs in segment 4, where the MT reverts the meaning of "There is little evidence..." to "Ni malo dokazov..."; again all three annotators agree in the severity level but not in the category.

5.3. Comparing both evaluation methods

While the Fluency/Adequacy evaluation method gives little insight into the specific issues that may have been improved or aggravated from one MT model to another, it seems relatively consistent in the scoring of different models across domains. If we compare the Fluency/Adequacy scores obtained for each text translated by the marian-v5 model with the results of the error annotation, correlation is low. According to the former, the most adequate and fluent translation was that of the legal text, and the least of the karst text. According to the number of annotated errors and edits, karst was the best and legal the worst. (The number of errors in Figure 3 is normalized to allow for better visual comparison.)

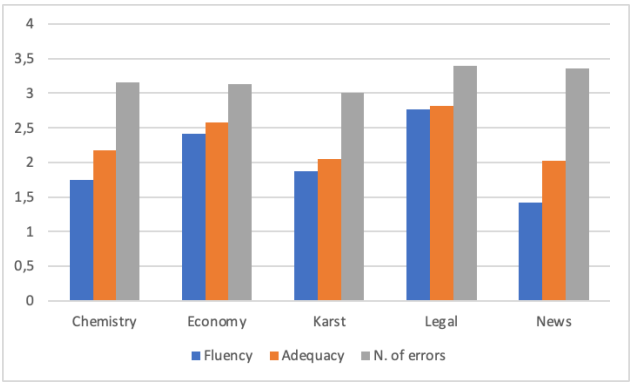

Figure 3: Comparing fluency, adequacy and number of errors per text.

6. Conclusion

We presented the results of human evaluation of MT using two well-known methodologies. The Fluency/Adequacy evaluation is relatively efficient and

fast, and the results are a useful indicator of the quality of different MT models. In general, the scores show high correlation with automatic metrics[3], with Nemo models achieving the highest automatic evaluation scores, followed by the Marian models and the baseline model, which is similar to what can be observed from the Adequacy/Fluency data. To measure the reliability of the Adequacy/Fluency ratings, we calculated the Cohen's kappa coefficient[4] for each document evaluated by a pair of evaluators. As somewhat expected, the agreement is fairly low with most of the values falling between 0.20 and 0.50. The fact that the evaluation was performed by students does not seem to significantly affect the results.

On the other hand, the evaluation through error annotation and post-editing requires a much higher level of effort, linguistic and extra-linguistic competence. Since each text was annotated by three students, a comparison of their decisions provides a valuable insight into the difficulty and subjectivity of the task. Agreement is low for all the parameters under observation: the number of errors marked, their categorization and their severity levels. Moreover, there is little correlation between the number of marked errors, their severity and the true quality of the machine translation. For the text which was the most specialized (Karst), contained a high number of un- or mistranslated terms and received the lowest Fluency/Adequacy score, the number of marked errors was the lowest of all. Student annotators with little or no expert knowledge of the domain will therefore find it difficult to correctly identify terminology errors, assess their severity or post-edit the text to a more accurate version.

Conversely, possibly owing to the fact that students of translation are still in the process of acquiring their language competence and are constantly reminded of the grammatical aspect of the texts they produce, their sensitivity to fluency-related issues is high, hence linguistic and stylistic errors are still often perceived as major. This might explain why the two texts which were most accessible and easy to understand received the highest number of marked errors.

In retrospect, the postediting and error annotation task was too difficult for advanced students of translation and failed to provide meaningful insights into MT quality, for several reasons: Firstly, the texts were too specialized and difficult to understand for non-experts. While students were free to use all available resources, some of the terminological expressions would require extensive research to resolve and the students lacked the time, motivation or skill to perform such research. Secondly, to ensure higher agreement in the severity and category of errors, students should have received training, a test run and much more comprehensive annotation guidelines with English-Slovene examples. Finally, the annotation environment in MemoQ with the rather fine-grained MQM/DSDE error typology is cumbersome and unintuitive, which probably affected the results.

We nevertheless believe that the experiments were valuable both for researchers and annotators. As researchers in MT development and evaluation we have gained experience which will allow us to better design evaluation runs, select texts and train annotators, and the student annotators have been subjected to translation quality assessment and postediting, both of which are tasks frequently encountered in professional translation.

7. Acknowledgments

The project Development of Slovene in a Digital Environment (Slovene: Razvoj slovenščine v digitalnem okolju, RSDO) is co-financed by the Republic of Slovenia and the European Union under the European Regional Development Fund. The operation is carried out under the Operational Programme for the Implementation of the EU Cohesion Policy 2014–2020.

The authors thank the students of MA Translation at the Faculty of Arts, University of Ljubljana, for their participation in the task.

8. References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72. Association for Computational Linguistics.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016. Ten years of WMT evaluation campaigns: Lessons learnt. In: *Proceedings of the LREC 2016 Workshop Translation Evaluation–From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 27–34.

Jennifer Doyon, Kathryn B. Taylor, and John S. White. 1999. Task-based evaluation for machine translation. In: *Proceedings of Machine Translation Summit VII*, pages 574–578.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics* 108, no. 1 (2017), pages 121–132.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, and Jonathan M. Cohen. 2019. Nemo: a toolkit for building AI applications using neural modules. arXiv:1909.09577.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció* 12, pages 455–463.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In: *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Katrin Marheinecke. 2016. Can Quality Metrics Become the Drivers of Machine Translation Uptake? An Industry Perspective. *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 71–76.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael

[3] Automatic metric scores can be found at https://slobench.cjvt.si/

[4] Using the *cohen_kappa_score* function from the *sklearn* Python library.

Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. arXiv:1904.01038.

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.

Rico Sennrich, Antonio Valerio Miceli Barone, Joss Moorkens, Sheila Castilho, Andy Way, Federico Gaspari, Valia Kordoni, Markus Egg, Maja Popovic, Yota Georgakopoulou, Maria Gialama, Menno van Zaanen. 2017. TraMOOC—translation for massive open online courses: recent developments in machine translation. In: *20th Annual Conference of the European Association for Machine Translation*, EAMT.

Dimitar Shterionov, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'dowd, and Andy Way. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation* 32, no. 3, pages 217–235.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

John S. White, Theresa A. O'Connell, and Francis E. O'Mara. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In: *Proceedings of the First Conference of the Association for Machine Translation in the Americas*.