On reducing translation shifts in translations intended for MT evaluation

Maja Popović
ADAPT Centre
Dublin City University
Ireland
`maja.popovic@adaptcentre.ie`

Abstract

Automatic evaluation of machine translation (MT) is based on the idea that the quality of the MT output is better if it is more similar to human translation (HT). Whereas automatic metrics based on this similarity idea enable fast and large-scale evaluation of MT progress and therefore are widely used, they have certain limitations. One is the fact that the automatic metrics are not able to recognise acceptable differences between MT and HT. The frequent cause of these differences are translation shifts, the optional departures from theoretical formal correspondence between source and target language units for the sake of adapting the text to the norms and conventions of the target language. This work is based on the author's own translation experience related to the evaluation of MT output compared to the experience unrelated to MT. The main observation is that, although without any instructions in this direction, fewer translation shifts were performed than when translating for other purposes. This finding will hopefully initialise further systematic research both from the aspect of MT as well as from the aspect of translation studies (TS) and bring translation theory and MT closer together.

1 Introduction and Motivation

The notion of translation shifts (Catford, 1965; van Leuven-Zwart, 1989; van Leuven-Zwart, 1990;

Curys, 2006) is an important concept in translation theory. A shift has occurred if the translation procedure has been "oblique" instead of "direct/literal" (Vinay and Darbelnet, 1958) so that there are "departures from formal correspondence" (Catford, 1965). A translated text in a target language can differ from the original text in the source language in many aspects and levels (such as lexical, structural, discourse, etc.) and still be perfectly acceptable. Apart from the transformations necessary for the sake of grammatical well-formedness, it is common practice to introduce optional changes to adapt the text to the norms and conventions of the target language depending on the genre, domain, register, etc. Although such changes are not strictly necessary, the professional translators are expected to deliver texts which appear natural in the target language.

These shifts, although natural and necessary in human translation (HT), pose challenges if the translations are used for automatic evaluation of machine translation (MT) systems. The vast majority of MT automatic evaluation metrics is based on similarity between MT and HT, and availability of a heterogeneous set of human reference translations is largely beneficial for metrics' performance (Albrecht and Hwa, 2008; Popović et al., 2016b). In practice, however, only one reference HT is available and its characteristics can strongly affect the results of automatic evaluation.

A lot of work has been done exploring differences between different types of texts, such as between texts originally written in a given language and texts translated into a given language ("translationese") (Baroni and Bernardini, 2006; Rabinovich and Wintner, 2015; Wintner, 2016; Daems et al., 2017), human and machine translations (Ahrenberg, 2017), as well as post-edited MT

outputs (PEs) as a special case of HT (Čulo and Nitzke, 2016; Daems et al., 2017; Farrell, 2018). Overall, the main findings are that HTs differ from original language texts because the source language seems to always "leave a trace" in the translation ("shine through"); similarly, PE, although generally capable of reaching the same quality as HT, always carries a "trace" of the used MT system.

Less work can be found about relations between these facts and the evaluation of machine translation. Popović et al. (2016b) compared the use of PEs and HTs, and suggest that PEs should be used carefully for MT evaluation due to the bias of each PE towards its MT system caused by the previously mentioned system's "shining through". Two methods for reducing translation shifts in reference HTs in order to alleviate automatic MT evaluation are proposed by Ahrenberg (2006) and Fomicheva et al. (2015). Ahrenberg (2006) proposes a method to identify "simpler" HTs which have certain desirable properties and are not too complex for MT systems. Fomicheva et al. (2015) investigate rule-based paraphrasing methods to reduce shifts in HTs and generate additional reference translations. Nevertheless, there is still a lot of room for systematic and extensive experiments dealing with different HTs and their relation to the MT evaluation. For example, even though a large number of HTs intended for MT evaluation have been generated in the framework of the WMT shared tasks[1] (Bojar et al., 2018; Bojar et al., 2017) running since 2006 until the present, no information can be found about the translators, such as how many segments did each individual translate, what are their qualifications, what translation experience or credentials do they have, how are they linked to the domain(s) of the data sets, whether they were aware of the purpose of their work and whether it had any influence.

This work reports the first qualitative feedback related to this topic. The work is based on the author's translation experience related to the evaluation of MT output compared to the experience unrelated to MT. The MT-related experience covers both translation as well as post-editing of MT output in order to provide reference human translations for automatic MT evaluation and/or error analysis. The main observation is that the transla-

tion shifts tend to be diminished when translating for MT evaluation purposes thus producing less creative and more literal translations.

This finding will hopefully become an attractive direction for future work both for MT as well as for translation studies (TS) and hopefully eventually contribute to bringing translation theory and MT closer together.

2 Translator's Background and Experience

The author's educational background is technical and not related to translation studies. Nevertheless, she has been interested in languages and translation since childhood (probably as a result of being raised by a translator mother), and has been translating on a more or less regular basis.

Her experience unrelated to MT involves mainly translation of scientific technical texts, as well as correspondences and short summaries of literary texts, between Serbian (native language) and English or French (professional proficiency). It also includes translation between German and English (professional proficiency) of lecture scripts about Machine Learning, Pattern Recognition and Signal Processing for courses held at the RWTH Aachen University.

The author has extensive experience in human and automatic MT evaluation and error analysis, including assigning adequacy and/or fluency (related) scores, error annotation on different levels of granularity, developing automatic evaluation metrics, developing automatic error classification tools, as well as comparing automatic metrics and tools with the corresponding results of human evaluation and annotation.

Her translation and post-editing (PE) experience related to MT evaluation involves mostly the Serbian language (native proficiency) as well as Croatian (near-native proficiency) as target languages, and English, Spanish, German (professional proficiency) and Slovenian (advanced passive competence) as source languages. The translated/post-edited texts include movie subtitles, user reviews (mainly about movies), and news.

For all these tasks, the goal of the work was clear, however no external instructions were specified (such as minimal post-editing or keeping maximal similarity to the source text), but were independently defined by the author. For example, the PE2rr corpus (Popović and Arčan, 2016) con-

---

[1] http://statmt.org/wmt19/, http://www.statmt.org/wmt18/, etc.

tains post-edited MT outputs together with error annotations and was created mainly in order to enable assessment of automatic error classification tools. Some parts of this corpus (English-to-Serbian and German-to-Serbian MT outputs) were post-edited by the author, who decided to perform relatively light post-editing without taking into account preferences concerning style, lexical choice, grammatical structure, etc. The reasoning behind this decision was that state-of-the-art automatic error classification tools are certainly not able to distinguish such subtle error categories. On the other hand, some other parts of this corpus were taken from the TaraXÜ corpus (Avramidis et al., 2014) which was developed independetly and post-edited by other translators without stating any particular goal. Many of these PEs contain much more substantial changes, including separating one sentence into two or joining two sentences. The same PE guidelines were applied for identifying the main problems for MT between the closely related Croatian, Serbian and Slovenian languages (Popović et al., 2016a), where the PEs are used as reference translations. The author has also generated reference HTs for the IMDb corpus in order to enable fast development of the first English-to-Serbian MT system for IMDb reviews through the use of automatic evaluation metrics (Lohar et al., 2019). During this translation process, the author noticed that a number of translations could feel much more natural if they diverged from the close (literal) translations, however she abstained from introducing these shifts knowing that the final goal of the translation was evaluating an MT system.

An important fact is that none of the MT evaluations included comparison between MT quality and HT quality in order to estimate the remaining gap (Toral and Way, 2018), or to claim "human parity" (Hassan et al., 2018) (reassessed by Toral et al. (2018)) "cracking NMT"[2] or similar.

3 Observations on translation shifts

The main observation about both translation and post-editing processes is a tendency towards a balance between two antagonised aspects: maximal similarity between source and translated texts and naturalness of the generated text in the target language. On the one hand, paraphrasing the close version and shifting away from the source is nor-

mal and natural, and the most common in practice. On the other hand, keeping the close version would ensure more reliable MT evaluation results, as suggested in (Ahrenberg, 2006; Fomicheva et al., 2015). Using a close HT version as a reference enables better and easier identification of potential drawbacks of the system related to the mandatory changes due to systemic differences between the languages. Therefore the author performed less paraphrasing and fewer shifts than usual (when the purpose of translation was not related to MT evaluation). The generated translations were thus often stylistically flawed but grammatically correct and accurate.

The observed divergences between the original texts and translations can be divided in three groups. First group of divergences was completely avoided:

- merging or splitting original sentences

  Sentences within a paragraph could be sometimes organised differently than in the source language text (merging two sentences, splitting one sentence into two, or even reorganising some parts); taking into account that state-of-the-art[3] MT systems as well as automatic evaluation metrics work only on the sentence (segment, line) level, such changes would heavily affect the MT evaluation.

- adding extra information on a sentence level

  Although in some cases adding extra information to a traslated sentence contributes to the naturalness of the generated text, the translator abstained from it because it is difficult for the MT systems to generate such content properly.

The second group of shifts was applied, however less frequently than usual. These shifts were mainly introduced when a closer version, although grammatically correct, would require awkward constructions. Some of these shifts were applied more frequently than others, and the following list is roughly ranked according to the increasing frequency:

- changing passive voice into active voice

  The passive voice (which is very frequently used in English, for example "the criminal

[3]Some of the on-going MT research attempts to go beyond the sentence level.

was arrested by the police") was kept in the translated text whenever possible.

- replacing a pronoun with the corresponding noun

  This shift is to some extent similar to the maximally avoided "adding extra information". Nevertheless, in some cases the usage of a pronoun instead of the noun can lead to ungrammatical sentences, so this type of divergence was introduced in such cases (this can happen with the English pronoun "it" which can be rather ambiguous and often cannot be translated directly).

- changing negation structure

  Different languages have different negation rules, and preservation of the source language structure is not always possible. However, possible stylistic changes were avoided (for example, "it is not necessary" can be replaced by "it is unnecessary").

- choosing a less common lexical option

  The usage of less frequent synonyms can add subtlety to the translated text and improve its naturalness. However, the author chose a more common option in the majority of such cases.

- changing phrase order

  Whenever possible, the order of phrases in the source text was preserved in the translation (this happens often in free-word-order target languages).

The third group of shifts was not avoided:

- omitting/using pronoun in pro-drop target languages

  Pro-drop languages do not always require a personal pronoun because this information is encoded in the verb morphology. When translating into a pro-drop language, the pronouns were always used naturally, independently of the pronouns in the source language.

- changing verb tense, aspect or mood

  Different languages can have different usage of verb tense, aspect and mood. The verbs were always used naturally in the translations, independently of the source language.

4 Discussion and future research directions

The reported translator's qualitative feedback indicates that the HTs specifically dedicated to the MT evaluation might have different properties than HTs generated for other purposes, similarly to differences between other types of texts (HTs and original texts, PEs and HTs). These differences may be reflected by different types and amounts of introduced translation shifts.

The reported observations open a number of questions and directions for future work. Having been written by an author who is mainly involved in MT evaluation, it is possible that some important points from TS have been missed in this paper. Nevertheless, the main goal of this work is to give an account of a potential common ground of translation procedures and MT, as well as to initialise further research on the topic, both from the MT evaluation point of view as well as from the aspect of TS.

4.1 Systematic analysis of translation shifts

First of all, it would be interesting to see the results of a systematic quantitative analysis, such as extracting statistics of different shifts in different HTs. The potential of automatic differentiation between HTs could be investigated, too.

Furthermore, several parameters should be taken into account in the future. One factor is the language pair (and translation direction), because each language pair involves distinct sets of mandatory and optional shifts. Another important factor which can have influence on the amount of performed/avoided translation divergences is translators' background, including his/her attitude towards MT. Last but not least, the primary goal of the intended MT evaluation has to be clear and well defined. For the evaluation tasks described here, where the goals of the evaluation were to estimate the MT system's ability to generate accurate and well-formed text and/or to estimate the progress during development of an MT system, HTs with a reduced number of translation shifts are definitely more convenient. On the other hand, if a high quality of MT output is desired, evaluating on natural HTs with a rich lexical variety and a number of translation shifts is a better option. And if the aim of the evaluation is to explore the remaining gap between MT and HT, the used HTs should definitely be completely natural, containing

a number of shifts, having a large lexical variety, and should originate from demanding genres and domains. Ideally, these HTs should be free from any influence of MT. An objective and informative analysis in this direction was carried out on literary texts (Toral and Way, 2018).

4.2 Bringing together translation procedures and MT

In the process of translating, translators transform one linguistic system which is given (in the source language) to another which they partially can adapt to their preferences (in the target language). This transformation can be performed using different translation procedures, ranging from direct word-to-word mapping to complex adaptations. Broadly speaking, translators can choose one of the two methods, namely literal (direct) translation (staying as faithful as possible to the source) and oblique (complex) translation (adapt the source language text to the target language) (Vinay and Darbelnet, 1958). This has also been referred to as literal in contrast to natural translation (Newmark, 1988), or, for literary texts translated into English, domesticating the text (bringing the text to the reader in the target language), in contrast to foreignizing the text (bringing the reader to the text similar to the source language) (Venuti, 2001).

The question of which kind of translating procedure is better has been the focus of a discussion for a very long time. Vinay and Darbelnet (1958) claim that the literalness should be preserved as much as possible and the oblique translation methods should only be used with good reason and within strictly defined limits. A similar position is presented by Newmark (1988), stating that the only valid argument against an acceptable literal translation is if it seems unnatural or clumsy in the target language. Venuti (2001) argues that "foreignising" (literal translation) is appropriate for literary translation in order to maximally preserve the original linguistic effect, whereas "domesticating" should be implemented in technical translation in order to ensure immediate intelligibility.

With the emergence of MT technologies, the (positive and negative aspects of) literality might be revisited including the additional MT point of view. Several general strategies for approaching MT from the perspectives of TS were proposed by Čulo (2014), although connecting translation procedures and MT has not been mentioned. A

step in this direction is described in (Jones and Irvine, 2013), where the authors investigate potentials and limits of statistical MT to perform literal vs. oblique translation. Exploring the new state-of-the-art MT approach, namely neural machine translation (NMT), in this sense would be a very interesting line of research, especially taking into account the general ability of NMT systems to produce fluent translations. It can be supposed that some of the free/flexible/oblique translations would be easier for an (N)MT system to generate than some others. A systematic analysis of translation shifts, possibly including different types of text (for example scientific and literary) and more than one NMT system, would certainly bring interesting insights.

Another line of future work is adding the MT evaluation aspect into the debate about literal vs. oblique translation. While this discussion may appear rather philosophical at first sight, it is connected with some important practical aspects, such as the previously mentioned final goal of the intended MT evaluation (obtaining an overall numeric score, ranking two or more MT outputs, analysis of grammatical errors, analysis of lexical and/or stylistic errors, comparing MT with HT, etc.).

Furthermore, MT evaluation can be interpreted as the purpose of the translated text in the framework of Skopos theory (Reiss and Vermeer, 1984). Skopos theory is another translation perspective which is no longer limited by conventional source-text orientated views. It focusses on a purpose of a translated text, and this purpose then determines translation strategies and procedures. In order for the translator to be able to interpret the purpose and apply appropriate strategies for this purpose, a translation brief provided by the client is deemed necessary (Nord, 2006). This brief should contain information about the intended function of the target text, the target text recipient, the time and place of text reception, the medium over which the text will be transmitted, and the motive for the production or reception of the text. For translations intended for MT evaluation, such brief would have MT evaluation as purpose, MT researchers and developers as recipients, assessment and development of MT system(s) as the motive for the production/reception of the text.

5 Conclusions

This work reports a qualitative feedback from the author, a translator with an extensive experience in MT evaluation who has tended to perform fewer translation shifts than usual when working on a task related to MT evaluation. The divergences which were thoroughly avoided are related to keeping the similarity on the sentence level, since state-of-the-art MT systems still cannot go beyond this level. Other lexical and structural divergences were diminished due to the awareness of the translation purpose.

In spite of this experience, the author is not convinced that introducing such guidelines (for translation shifts and/or broader ones) on a more general level would be beneficial for MT research. One reason are the previously mentioned claims and hype about MT reaching human parity. Another reason is the objectively rapid development and improvement of MT systems. If more literal/less natural translations would be provided and used for MT on a large scale, some of these data sets could be repeatedly used for a longer period of time (for example, the English-German test set from WMT 2014 is still widely used for assessing new MT systems in 2019). In this way, the overall progress of MT would be measured using too literal and not fully natural texts, which does not seem appropriate.

On the other hand, a number of MT systems is still being developed under sub-optimal conditions (such as low-resourced languages, low computational resources, etc.). For such systems, guidelines on reducing translation shifts would help to better identify main problems and directions for further development. One possibility would be to specify the goal of each particular evaluation as purpose in a translation brief, for example "finding most prominent errors in MT hypotheses", "estimating post-editing effort", "measuring the progress of an MT system over time", etc.

The presented observations and suggestions are certainly influenced by the perspective of the author whose experience is closely related to MT evaluation. Still, these findings will hopefully initialise future research on the topic and bring translation procedures and MT together. This line of research could generally bring the fields of MT and TS closer, and initialise more collaborations between translators and MT researchers.

References

Ahrenberg, Lars. 2006. Codified close translation as a standard for MT. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT 2005)*, pages 218–229, Budapest, Hungary, May.

Ahrenberg, Lars. 2017. Comparing machine translation and human translation: A case study. In *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*, pages 21–28, Varna, Bulgaria, September.

Albrecht, Joshua S. and Rebecca Hwa. 2008. Regression for machine translation evaluation at the sentence level. *Machine Translation*, 22(1-2):1–27, March.

Avramidis, Eleftherios, Aljoscha Burchardt, Sabine Hunsicker, Maja Popović, Cindy Tscherwinka, David Vilar Torres, and Hans Uszkoreit. 2014. The taraXÜ Corpus of Human-Annotated Machine Translations. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2679–2682, Reykjavik, Iceland, May.

Baroni, Marco and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, September.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark, September.

Bojar, Ondej, 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, Volume 2: Shared Task Papers*,

pages 272–307, Belgium, Brussels, October. Association for Computational Linguistics.

Catford, John Cunnison. 1965. *A Linguistic Theory of Translation: An Essay in Applied Linguistics*. Oxford University Press, Oxford, UK.

Čulo, Oliver and Jean Nitzke. 2016. Patterns of terminological variation in post-editing and of cognate use in machine translation in contrast to human translation. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT 2016)*, pages 106–114, Riga, Latvia.

Čulo, Oliver. 2014. Approaching machine translation from translation studies – a perspective on commonalities, potentials, differences. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT 2014)*, pages 199–206, Dubrovnik, Croatia, June.

Curys, Lea. 2006. Building a resource for studying translation shifts. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1240–1245, Genoa, Italy, May.

Daems, Joke, Orphée De Clercq, and Lieve Macken. 2017. Translationese and post-editese : how comparable is comparable quality? *Linguistica Antverpiensia New Series – Themes in Translation Studie*, 16:89–103.

Farrell, Michael. 2018. Machine translation markers in post-edited machine translation output. In *Proceedings of the 40th Conference Translating and the Computer*, pages 50–59, London, UK, November.

Fomicheva, Marina, Núria Bel, and Iria da Cunha. 2015. Neutralizing the effect of translation shifts on automatic machine translation evaluation. *Computational linguistics and Intelligent Text Processing*, pages 596–607.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, , and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. In *arXiv*.

Jones, Ruth and Ann Irvine. 2013. The (Un)faithful Machine Translator. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–101, Sofia, Bulgaria, August.

Lohar, Pintu, Maja Popovi'c, and Andy Way. 2019. Building English-to-Serbian Machine Translation System for IMDb Movie Reviews. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2019)*, Florence, Italy, August.

Newmark, Peter. 1988. *A Textbook of Translation*. Prentice Hall, New York.

Nord, Christiane. 2006. Translating as a purposeful activity: A prospective approach. *TEFLIN*, 17(2):131–143.

Popović, Maja and Mihael Arčan. 2016. PE2rr Corpus: Manual Error Annotation of Automatically Preannotated MT Post-edits. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portoro, Slovenia, may.

Popović, Maja, Mihael Arcan, and Filip Klubička. 2016a. Language Related Issues for Machine Translation between Closely Related South Slavic Languages. In *Proceedings of the 3rd Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 43–52, Osaka, Japan, December.

Popović, Maja, Mihael Arçan, and Arle Lommel. 2016b. Potential and limits of using post-edits as reference translations for MT evaluation. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT 2016)*, pages 218–229, Riga, Latvia, June.

Rabinovich, Ella and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432, December.

Reiss, Katharina and Hans Vermeer. 1984. *Towards a General Theory of Translational Action; Skopos Theory Explained (Grundlegung einer allgemeinen Translationstheorie)*. Routledge, London.

Toral, Antonio and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment. Machine Translation: Technologies and Applications*, pages 263–287. Springer.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the 3rd Conference on Machine Translation (WMT 2018)*, pages 113–123, Belgium, Brussels, October. Association for Computational Linguistics.

van Leuven-Zwart, Kitty M. 1989. Translation and original: similarities and dissimilarities, i. *Target*, 1(2):151–181.

van Leuven-Zwart, Kitty M. 1990. Translation and original: similarities and dissimilarities, ii. *Target*, 1(2):69–95.

Venuti, Lawrence. 2001. Strategies of translation. In Baker, M., editor, *Routledge Encyclopedia of Translation Studies*, pages 240–244. Routledge, London & New York.

Vinay, Jean-Paul and Jean Darbelnet. 1958. *Comparative Stylistics of French and English: A Methodology for Translation*. John Benjamins.

Wintner, Shuly. 2016. Translationese: Between human and machine translation. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016): Tutorial Abstracts*, pages 18–19, Osaka, Japan, December.