Evaluation of adaptive machine translation from a gender-neutral language perspective

Aida Kostikova
Ghent University
aida.kostikova@ugent.be

Todor Lazarov
Ghent University
New Bulgarian University
tdlazarov@gmail.com

Joke Daems
joke.daems@ugent.be

Abstract

In this paper, we attempt to analyse the problem of conveying gender-neutral language when working with notional and grammatical languages (English and German) from the point of view of adaptive machine translation (MT). More specifically, we assess the efficiency of adaptive MT when it comes to gender-neutral language use, the purpose of which is to "reduce gender stereotyping, promote social change and contribute to achieving gender equality". We conclude that the initial output largely reflects cases of misgendering and generic masculine – problems that are well documented in the MT field, but which still remain unresolved. Moreover, our experiment revealed that ModernMT faces systematic difficulties in adapting to gender-neutral language when working with the English-German translation direction.

Introduction and Related Work

As the adoption of gender-neutral language (GNL) becomes more widespread, it is increasingly important to consider how these trends can be reflected in natural language processing (NLP) applications, especially given the fact that the purpose of GNL is to "reduce gender stereotyping, promote social change and contribute to achieving gender equality" (Papadimoulis, 2018: 3). To date the task of reflecting such linguistic trends as GNL has been addressed within the field of uncustomised, generic machine translation (MT) (Dev et al., 2021; Prates et al., 2019). At the same time, there are other promising and efficient solutions with the capacity of being more flexible in terms of use of gender-fair language. For example, adaptive MT is a technology which is characterised by its ability to learn from its users, make suggestions and improve accuracy over time. Adaptive MT builds on the concept of human-in-the-loop learning, which is the process by which a machine learning model receives and utilizes human intervention or feedback (Finkelstein, 2020).

Moreover, while notional gender languages, such as English, are more or less consistent in GNL strategies, more morphologically rich languages present a challenge in terms of adapting a universal gender-fair approach (Stahlberg et al., 2007). Existing strategies in German, for example, include declension rules modifications, various

gender-neutral wordings, neopronouns (Hornscheidt and Sammla 2021), and in most cases represent an individual, rather than systematic linguistic choice. This fits the purpose of adaptive MT, which adjusts to personal linguistic preferences, which can also include GNL use.

In this paper, we will assess the efficiency of adaptive MT when it comes to GNL use, focusing on non-binary oriented language use (that is, language that avoids bias toward not only females, but also individuals who identify outside the gender binary) (del Rio-Gonzalez, 2021). In particular, we will be putting the ModernMT[1] engine to the test and analyse whether and to which degree it can be retrained "on-the-fly" in attempting to ensure gender-neutrality in translation. English-German was chosen as a main working language pair in order to analyse how adaptive model of the engine adjusts the output to complex GNL modifications specific for grammatical gender languages as German (Stahlberg et al., 2007).

Methodology

In order to achieve the objectives of the study, we translated a text with the help of adaptive MT, identified bias which might be reflected in the initial output, concentrating exclusively on bias leading to under-representation of certain groups (Savoldi et al., 2021) and evaluated the adaptive model of the engine by post-editing the MT output and registering the process with the help of CharacTER (translation edit rate on character level) (Wang et al. 2016) and KSR (keystroke ratio), which was registered with the help of Inputlog, a keystroke logging program. Only gender-related items were edited. ModernMT, an adaptive MT system (integrated in MateCat, an online computer assisted translation tool), was chosen as the basis for the study. Its distinctive feature is that no changes are reflected in its base engine, and all modifications are introduced with the help of an "instance-based adaptive NMT" technology, which means that a system's generic model incrementally updates with the help of the dynamic configuration of the learning algorithm's hyperparameters (Farajian et al., 2017).

Texts developed by the International Quidditch Association[2] were used as the material for the study, as their texts are written in GNL and are available in different languages. The text size was 1138 words (divided into 45 segments) and it included 29 examples of gender-ambiguous nouns and a gender-neutral pronoun *they* in its different inflected forms, and we also made sure every word occurred at least three times in the text to increase the likelihood of the system being able to adapt after two repetitions. As a first step, an initial output generated by a baseline system was evaluated against a group of linguistic criteria derived from the European Parliament's guide on GNL. Then, the output was edited using the adaptive function of the ModernMT engine, with the increased emphasis on GNL forms, not on the overall quality of translation. As existing strategies in German are very complex due to the morphologically rich grammatical gender system (Hornscheidt and Sammla 2021), and represent an

[1] https://modernmt.com

[2] https://iqasport.org

individual, rather than systematic linguistic choice, two approaches were chosen to test the performance of an engine when working with potentially challenging elements: De-E-System, which introduces a whole new system of declension rules and neopronouns: for example, in order to eliminate a masculine gender marker in the plural noun *Spieler* (pl. *players*), which is used to refer to a group of people whose gender is unknown or irrelevant, it was changed to *Spielerne (Spielerne können in ihrem eigenen Namen mit den Offiziellen sprechen – players may speak to officials on their own behalf)*; and the gender star — a nonstandard typographic style, where an asterisk (*) is used to separate gendered inflections in the German language to include individuals who identify themselves outside of the gender binary, like in the word *Spieler*innen*: *Jedes Team besteht aus zwischen 7 und 21 Spieler*innen* (*Each team is made up of between 7 and 21 players*).

Results and Discussion

The first objective of our study was the manual evaluation of gender bias, which may be present in the initial output of the MT system. Two different trends were identified during the analysis: generic masculine and misgendering.

Baseline Model

29 of 29 nouns were always translated in the masculine form, and none of the sentences were translated with at least double gender names. For example, *speaking captain* was always translated as *der sprechende Kapitän*, *player – ein/der Spieler, coach – ein/der Trainer, the Chair – der Vorsitzende, the IQA CEO – der IQA CEO*. Nouns, articles and pronouns in the plural have also been translated on the basis of the masculine form: *Teammitarbeiter (team staff), Kapitäne (captains), Spieler (players)*, although the German language has means for avoiding using generic masculine in the plural, which, however, are limited to the binary gender system: for example, using feminine-masculine word pairs, using feminine-masculine word pairs (e.g., *Ingenieurinnen und Ingenieure – engineers*).

The reason for that could be that its baseline model is trained in the same way as generic MT systems (Farajian et al., 2017), which are prone to pre-existing bias – any asymmetries which are rooted in society at large or languages' structure and use (Silveira, 1980; Hamilton, 1991). If present in the training data, asymmetries in the semantics of language use and gender distribution are respectively inherited by the output of the MT (Caliskan et al. 2017).

Misgendering

Another problem identified in the first part of the experiment was misgendering, which describes cases where a person is addressed by a gendered term that does not match their gender identity. For example, *they* (with one instance of *them* and five instances of *their*), which was present in 16 segments, was not translated with a gender-neutral

term in any of the cases. As noted by Dev et al. (2021), language models are prone to misgendering when there is insufficient information to disambiguate the gender of an individual, and so they default to binary pronouns and binary-gendered terms, as we observe in the case of the baseline of ModernMT.

In most cases, the pronouns *they/them/their* were treated as plural pronouns in the third person even if there is a direct reference to a single person: *Die IQA soll den Beschwerdeführer darüber informieren, wann sie zusätzliche Mitteilungen erwarten können* (*The IQA should inform the complainant as to when they can expect additional communication*). Moreover, in some cases, *they* was translated as a masculine singular pronoun: *Wenn der Kapitän in das Spielfeld zurückkehrt, nimmt er die Rolle des Kapitäns wieder auf* (*If the captain returns to the pitch, they shall resume the role of the captain*). This is also in line with the observation made by Dev et al. (2021), who noted that language models can also misgender individuals even when their pronouns are provided.

These findings indicate that the text translated by the baseline system would require a considerable amount of post-editing. In the next section, we verify whether using the adaptive function of the engine reduces that post-editing effort.

4.2. Adaptive model

CharacTER, which is common in post-editing efforts studies (Bentivogli et al., 2016), was calculated for each segment, and its change during the translation process indicated the rate at which the system adapts to the edits: for example, 0 would mean that the segment did not need any post-editing, and increase in the number of such sentences by the second half of the text (starting from the segment 22) would indicate that the system started picking up the gender-neutral forms. Possible edits included the insertion, deletion, and correcting punctuation errors; shifts of word sequences were avoided where possible (Snover et al., 2006).

As TER-derived metrics heavily depend on the length of the sentence, the text was pre-processed to ensure that every segment in the text is of an average length (around 65 characters) and has a comparable number of potentially problematic items (for instance, nouns, pronouns, articles). CharacTER scores, which reflect the final results of translation process, were complemented by KSR, to measure the number of keystrokes (and, therefore, the actual editing process) that are needed to edit the MT output.

De-E-System

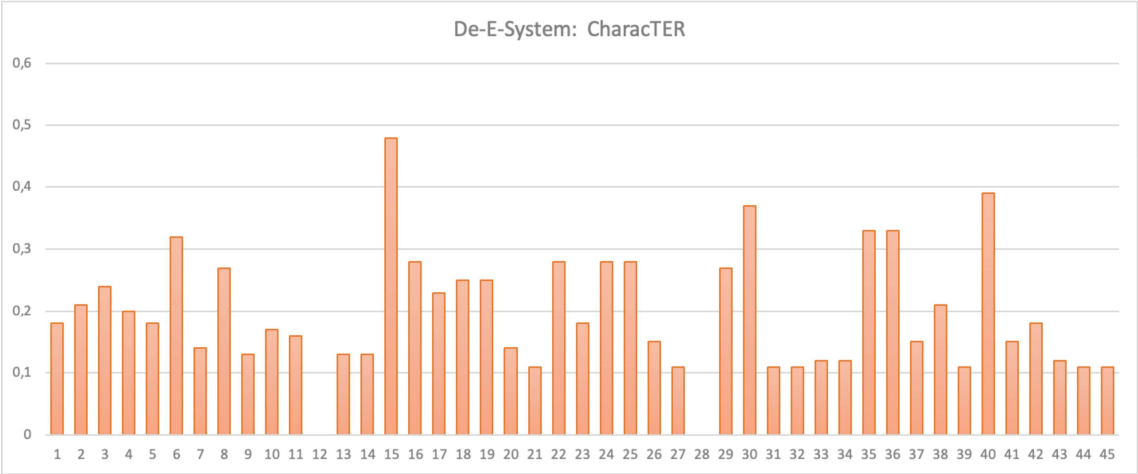

Figure 1: CharacTER scores for each segment edited according to the De-E-System

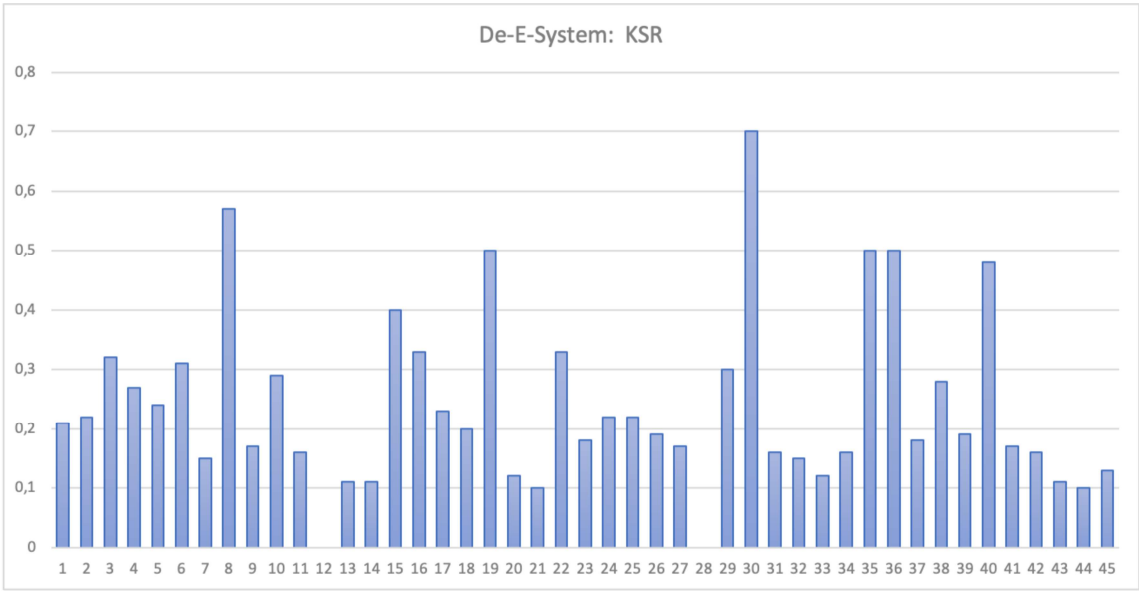

Figure 2: KSR for each segment edited according to the De-E-System

As can be seen from figures 1 and 2, only segments 12 and 28 have reached a zero value, which are in fact exact translation memory matches. The system did not show any improvements in adapting to the edits introduced by a translator; in fact, gender-ambiguous words invariably took a masculine form: for example, the word *a player* was translated as *ein* (or *der*) *Spieler* throughout the text after each segment was edited in line with the GNL strategies.

Gender Asterisk

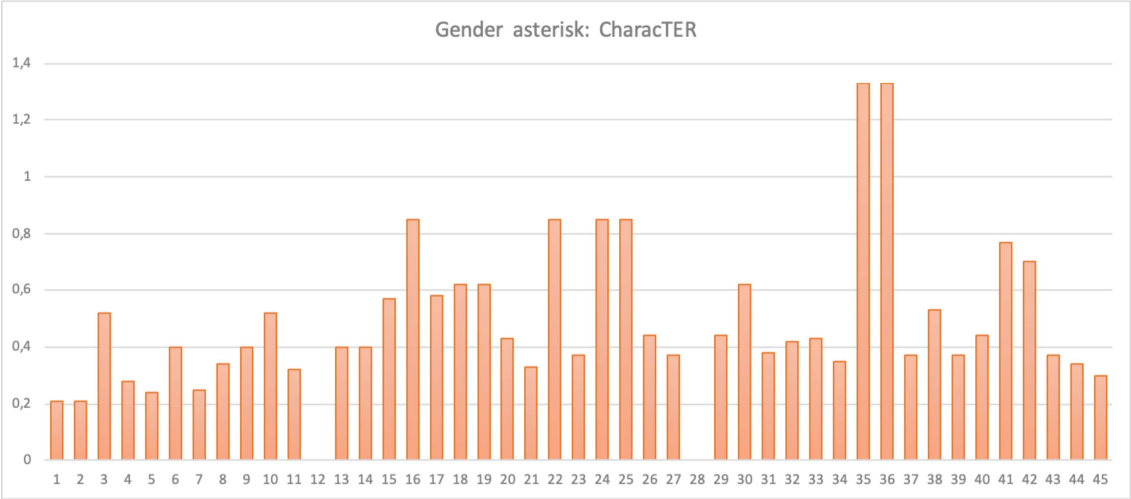

Figure 3: CharacTER scores for each segment edited according to the "gender asterisk" system

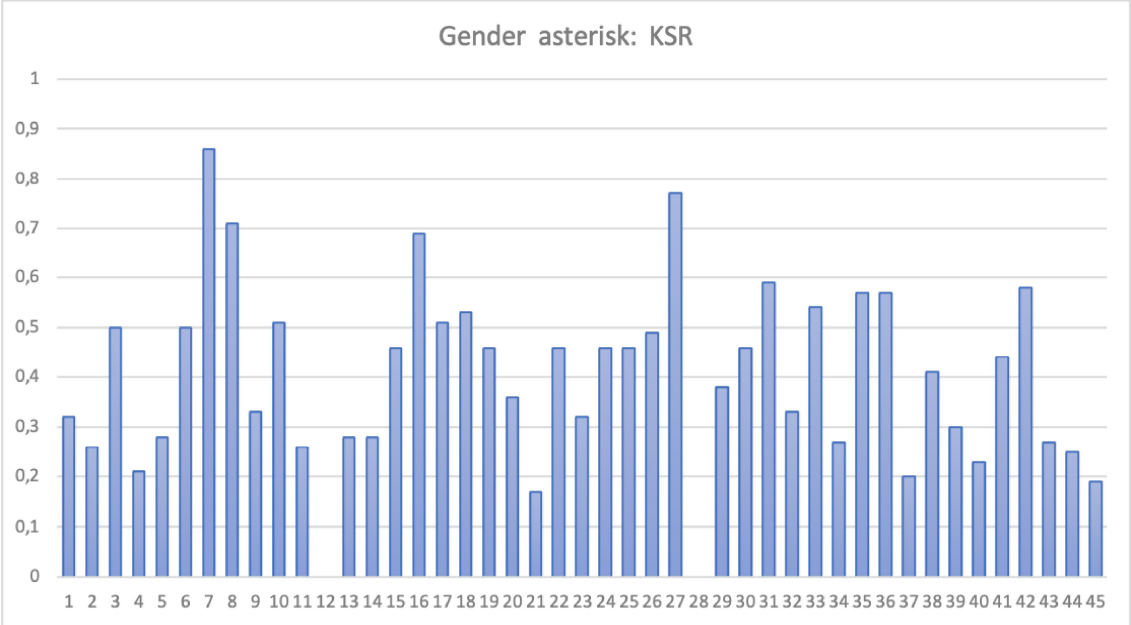

Figure 4: KSR for each segment edited according to the "gender asterisk" system

Similar results are observed with the gender asterisk: overall, this system requires more post-editing effort than the De-E-System, due to larger number of characters required to align the text with the gender-neutral strategies. Nevertheless, the system failed to adopt any changes made during the translation process, as no improvements are seen in CharacTER or KSR. It should also be noted that the active use of typographic characters did not have any effect on the rest of the text and no distortions were detected. On the other hand, for each edited segment the system reported a symbol mismatch, which

occurs when that the source and target segments do not contain the same elements and symbols. As in the case with the De-E-System, zero values are seen for segments 12 and 28, which were exact translation memory matches.

Conclusion and Future Work

In this paper, we analysed the problem of conveying GNL when working with the English-German translation direction from the point of view of adaptive MT. More specifically, we assessed the efficiency of adaptive MT by putting its baseline and adaptive functionality to the test. We conclude that the initial output largely reflects cases of misgendering and generic masculine – problems that are well documented in the MT field, but which still remain unresolved.

Some issues were also detected when working with the adaptive part of ModernMT: no progress in adaptation speed was registered when working with GNL, except for the cases of TM auto-propagation. For future work, we will additionally train the ModernMT engine by feeding it a translation memory containing GNL, and we will compare the adaptivity of another MT system, Lilt. A preliminary experiment with Lilt showed that this engine is capable of adapting to gender-neutral forms: for example, it suggested the gender-neutral noun Kapitän*in in the tenth segment.

References

Bentivogli, Luisa, Nicola Bertoldi, Mauro Cettolo, Marcello Federico, Matteo Negri, and Marco Turchi. 2016. On the evaluation of adaptive machine translation for human post-editing. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2), pages 388-399.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. In *Science, 356(6334)*, pages 183-186.

Farajian, M. Amin, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127-137.

Finkelstein, Paige. 2020. Human-assisted Neural Machine Translation: Harnessing Human Feedback for Machine Translation. University of Washington.

Hamilton, Mykol C. 1991. Masculine bias in the attribution of personhood: People= male, male= people. In: *Psychology of Women Quarterly,* 15(3), pages 393-402.

Hornscheidt, Lann and Sammla, Ja'n. 2021. *Wie schreibe ich divers? Wie spreche ich gendergerecht? Ein Praxis-Handbuch zu Gender und Sprache*. Insel Hiddensee: w_orten & meer.

María del Río-González, Ana. 2021. To Latinx or not to Latinx: a question of gender inclusivity versus gender neutrality. In *American Journal of Public Health, 111(6)*, pages 1018-1021.

Papadimoulis, Dimitros. 2018. *Gender-neutral language in the European Parliament*. Brussels: European Parliament.

Prates, Marcelo OR, Pedro H. Avelar, and Luís C. Lamb. 2019. Assessing gender bias in machine translation: a case study with Google Translate. In *Neural Comput & Applic* 32, pages 6363–6381.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. In *Transactions of the Association for Computational Linguistics; 9*, pages 845–874.

Silveira, Jeanette. 1980. Generic masculine words and thinking. In *Women's Studies International Quarterly,* 3(2-3), pp. 165-178.

Snover, Matthew, 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.

Stahlberg, Dagmar, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. In *Social communication*, pages 163-187.

Wang, Weiyue, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505-510.