Mitigating Gender Bias in Machine Translation with Target Gender Annotations

Artūrs Stafanovičs*†‡ and **Toms Bergmanis*†‡** and **Mārcis Pinnis†‡**

†Tilde / Vienības gatve 75A, Riga, Latvia
‡Faculty of Computing, University of Latvia / Raiņa bulv. 19, Riga, Latvia
`{firstname.lastname}@tilde.lv`

Abstract

When translating "*The secretary asked for details.*" to a language with grammatical gender, it might be necessary to determine the gender of the subject "*secretary*". If the sentence does not contain the necessary information, it is not always possible to disambiguate. In such cases, machine translation systems select the most common translation option, which often corresponds to the stereotypical translations, thus potentially exacerbating prejudice and marginalisation of certain groups and people. We argue that the information necessary for an adequate translation can not always be deduced from the sentence being translated or even might depend on external knowledge. Therefore, in this work, we propose to decouple the task of acquiring the necessary information from the task of learning to translate correctly when such information is available. To that end, we present a method for training machine translation systems to use word-level annotations containing information about subject's gender. To prepare training data, we annotate regular source language words with grammatical gender information of the corresponding target language words. Using such data to train machine translation systems reduces their reliance on gender stereotypes when information about the subject's gender is available. Our experiments on five language pairs show that this allows improving accuracy on the WinoMT test set by up to 25.8 percentage points.

1 Introduction

Most modern natural language processing (NLP) systems learn from natural language data. Findings of social sciences and corpus linguistics, however, indicate various forms of bias in the way humans use language (Coates, 1987; Butler, 1990; Fuertes-Olivera, 2007; Rickford, 2016). Thus the resulting NLP resources and systems also suffer from the same socially constructed biases, as well as inaccuracies and incompleteness (Jørgensen et al., 2015; Hovy and Søgaard, 2015; Prates et al., 2019; Vanmassenhove et al., 2019; Bordia and Bowman, 2019; Davidson et al., 2019; Tan and Celis, 2019). Due to the prevalent use of NLP systems, their susceptibility to social biases becomes an increasingly significant concern as NLP systems not only reflect the biases learned but also amplify and perpetuate them further (Hovy and Spruit, 2016; Crawford, 2017; HLEG, 2019).

This work concerns mitigating the manifestations of gender bias in the outputs of neural machine translation (NMT) systems in scenarios where the source language does not encode the information about gender that is required in the target language. An example is the translation of the English sentence "*The secretary asked for details.*" into Latvian. In English, the gender of "*secretary*" is ambiguous. In Latvian, however, there is a choice between the masculine noun "*sekretārs*" and the feminine noun "*sekretāre*". In cases when sentences do not contain the necessary information, NMT systems opt for translations which they have seen in training data most frequently. Acquiring the necessary information, however, might require analysis of the text beyond the level of individual sentences or require incorporation of external knowledge.

Falling back to biases, however, happens not only in the absence of the required information as NMT systems produce stereotyped translations even when clues about the subject's correct gender are present in the sentence (Stanovsky et al., 2019). This is in line with findings by Vanmassenhove et al. (2019) who suggest that NMT systems produce biased outputs not only because of the biases

---

*First authors with equal contribution.

present in data but also due to their tendency to exacerbate them.

To provide means for incorporation of external and explicit gender information, we propose a method for training NMT systems to use word-level gender annotations. To prepare training data, we project grammatical gender information of regular target language words onto the corresponding source language words. Albeit in some cases redundant, we expect that the grammatical gender information contains a useful learning signal that helps narrowing down the lexical choice of the correct target translation. As a result, the NMT system learns to rely on these annotations when and where they are available. In particular, in experiments on five language pairs, we show that the methods proposed here can be used in tandem with off-the-shelf co-reference resolution tools to improve accuracy on the WinoMT challenge set (Stanovsky et al., 2019) by up to 25.8 percentage points.

1.1 Related work

Recent recommendations for ethics guidelines for trustworthy AI recommend removing socially constructed biases at the source, the training data, prior to model training (HLEG, 2019). An example of work on debiasing training data is Zhao et al. (2018) where authors identified sentences containing animate nouns and changed their grammatical gender to the opposite. Zmigrod et al. (2019) take it further by ensuring that not only the animate nouns but also the rest of the sentence is reinflected from masculine to feminine (or vice-versa), thus preserving the morpho-syntactic agreement of the whole sentence. The applicability of this line of work is still to be established as reinflecting sentences with co-references or pairs of parallel sentences in NMT pose an additional challenge.

A different take on addressing gender biases in NMT outputs is the work on alternative generation: given a gender-ambiguous source sentence and its translation, provide an alternative translation using the opposite gender. Habash et al. (2019) approach this as a gender classification and reinflection task for target language sentences to address the first person singular cases when translating from English into Arabic. Bau et al. (2018) analyze trained NMT models to identify neurons that control various features, including gender information, that are used to generate the target sentence. In practice, however, such solutions are limited to simple source sentences where only one alternative in the target language is possible.

A complementary approach is addressing gender bias in NMT as a problem of domain mismatch. When translating TED talks, Michel and Neubig (2018) propose to adapt the NMT model for each speaker's attributes, thus also implicitly addressing previously poorly translated first-person singular cases. Saunders and Byrne (2020) describe methods for NMT model adaptation using a hand-crafted gender-balanced dataset and a translation re-scoring scheme based on the adapted models.

The closest line of work to ours is the work on the incorporation of external gender information in the NMT input. Elaraby et al. (2018) and Vanmassenhove et al. (2018) prepend training data sentences with speaker gender information to improve spoken language translation when translating into languages with grammatical gender. Moryossef et al. (2019) undertakes a similar approach at the inference time using phrases (e.g. *"she said:"*) that imply the speaker's gender. The methods proposed in this work differ from the previous work in terms of annotation granularity: we propose to use token level annotations, while the previous work used one annotation per sentence. As our training data annotations are solely based on grammatical gender, preparing them does not require any external gender information. Thus our approach is also simpler in terms of training data preparation compared to the previous work (Elaraby et al., 2018; Vanmassenhove et al., 2018).

Social Impact We propose methods to mitigate the manifestations of gender bias in the outputs of NMT. Specifically, these methods provide explicit means to incorporate information about subjects referential or social gender in NMT, thus reducing gender-based stereotyping when translating into languages which encode for grammatical gender in animate nouns. An example of a use case and a beneficiary group is the translation of occupational nouns into languages which mark gender and people for whom stereotypes of their profession do not align with their gender. While these methods can relieve gender-based representational harms by reducing stereotyped translations, they, unfortunately, provide no means for better representation of non-binary gender identities.

2 Methods

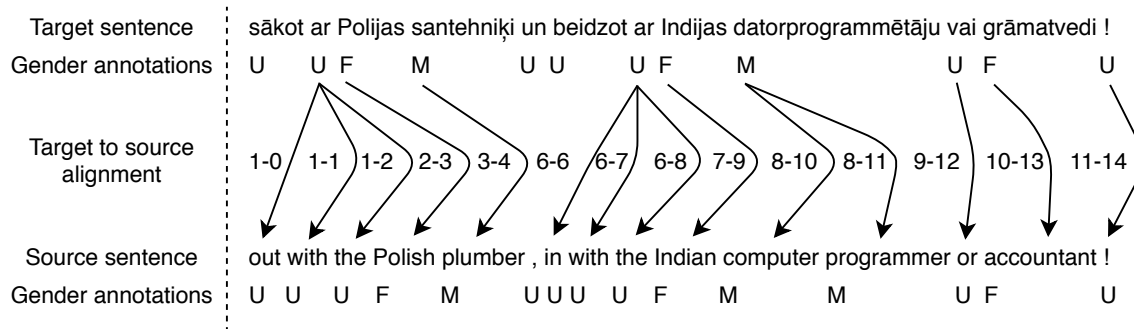When translating from languages without grammatical gender to languages with grammatical gender,

Figure 1: Illustration of target to source projections of grammatical gender annotations. Sample sentences taken from the English-Latvian development set of the WMT2017 News Translation Task.

certain words in the source sentence may not contain all the necessary information to produce an adequate and accurate translation. Examples are pronouns (e.g. *I, me, they, them, themselves*), animate nouns such as job titles and proper nouns such as names and surnames, which depending on the sentence context can be ambiguous and consequently can be translated poorly. Previous work has also shown that NMT systems are better at translating sentences that align with socially constructed gender stereotypes because they are more frequently seen in training data (Stanovsky et al., 2019; Prates et al., 2019).

To circumvent the degradation of NMT outputs due to 1) socially constructed biases and 2) absence of necessary information, we propose a method for training NMT systems to be aware of and use word-level target gender annotations (**TGA**). For training, we use data where regular source language words are annotated with the grammatical gender of their target language translations. We obtain such data by, first, morphologically tagging target language sentences to obtain information about their grammatical gender—F for feminine, M for masculine, N for neuter, and U for cases where grammatical gender is unavailable. Then, we use word-level statistical alignments to project this information from the target language to the source language words (see Figure 1 for an illustration). We use source-side factors (Sennrich and Haddow, 2016) to integrate the projected annotations as an additional input stream of the NMT system. To ensure that the NMT systems are capable of producing adequate translations when gender annotations are not available—a frequently expected case at the test time—we apply TGA dropout. We do so by randomly replacing annotations for a random number of words with U.

While useful for animate nouns, such annotations might seem otherwise redundant because the majority of nouns in training data can be expected to be inanimate. However, for some inanimate nouns, the target language grammatical gender annotations can help narrowing down the lexical choice during training. An example is the translation of *"injury"* into Latvian, where *"injury*|F" would result in *"trauma"* while *"injury*|M" would correspond to *"ievainojums"*. Besides disambiguating animate nouns, annotations also disambiguate the grammatical gender of pronouns, proper nouns. Furthermore, grammatical gender annotations also concern adjectives and verbs, which in some languages have to agree in gender with the nouns they describe. Consequently, we expect that during training the NMT model will learn to use these annotations, as they contain valuable information about words in the target sentence.

At inference time, we lean heavily on the observation that there the grammatical gender of animate nouns, pronouns, and proper nouns, and the intended referential gender coincide considerably. This is, however, a heuristic and not a rule (see Hellinger and Motschenbacher (2015) for counterexamples). Nevertheless, we assume that it is possible to use TGA in a referential sense of gender, thus injecting the NMT model with additional information about the subject's gender. Sources of such information can vary; in this paper, we showcase how to use TGA together with off-the-shelf co-reference resolution tools.

2.1 Evaluation: WinoMT Test Suite

To measure the extent to which gender annotations reduce NMT systems' reliance on gender stereotypes, we use the WinoMT test suite (Stanovsky et al., 2019). WinoMT builds on the previous work
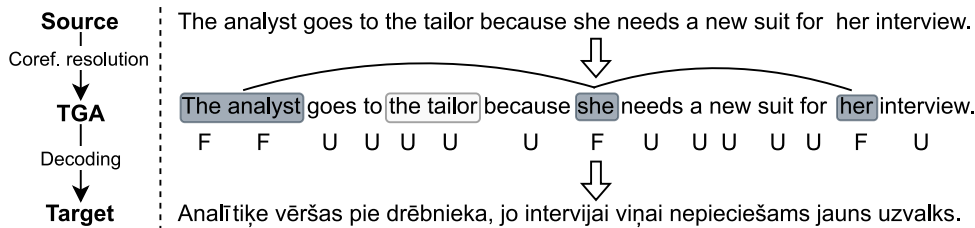
Figure 2: WinoMT test suite translation process with TGA distilled from the output of automatic coreference resolution tool.

on addressing gender bias in co-reference resolution by combining Winogender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018) datasets in a test suite for automatic evaluation of gender bias in MT. All sentences in the WinoMT test set follow the Winograd Schema where anaphora resolution is required to find an antecedent for an ambiguous pronoun (Hirst, 1981). In the case of datasets designed for evaluation of gender bias, the ambiguous pronoun refers to one of two entities which are referred to using titles of their professions. Professions and pronouns are chosen so that they either align with or diverge from the gender stereotypes of each profession as reported by the U.S. Bureau of Labor Statistics (Zhao et al., 2018).

WinoMT tests if the grammatical gender of the translation of an antecedent matches the gender of the pronoun in the original sentence. Testing is done by morphologically analysing the target translation and aligning it with the source sentence. The WinoMT test suite scores MT outputs using multiple metrics: **Accuracy** – the percentage of correctly translated antecedents, ΔG – difference in F_1 score between sentences with masculine and feminine antecedents, ΔS – difference in accuracy between the set of sentences that either align with or diverge from the gender stereotypes of each profession. Saunders and Byrne (2020) also propose to report **M:F** – ratio of translations using masculine and feminine antecedents.

3 Experimental Setting

Languages and Data In all our experiments, we choose one source language without grammatical gender and five Indo-European languages in which nouns have grammatical gender (see Table 1). For all language pairs, we use training data from WMT news translation tasks. We do the necessary cleaning and filtering with Moses (Koehn et al., 2007) pre-processing tools. To see how TGA is affected by data size, we also use much larger EN-LV propri-

	Source	# Sent.	News Test
EN-DE	WMT19	64.1M	2018
EN-FR	WMT15	39.1M	2015
EN-LV	Tilde	22.7M	2017
EN-LV	WMT17	4.5M	2017
EN-LT	WMT19	3.6M	2019
EN-RU	WMT17	25.0M	2015

Table 1: Training data set source and size in millions of sentences prior to adding TGA.

etary data that we obtain from Tilde Data Libarary by combining all EN-LV parallel corpora. The proprietary data are pre-processed using the Tilde MT platform (Pinnis et al., 2018). Table 1 summarizes training data source and size statistics prior to adding TGA. For all systems and language pairs, we use byte pair encoding (BPE) (Gage, 1994; Sennrich et al., 2016) to prepare joint source and target language BPE sub-word vocabularies. We use 30K BPE merge operations and use a vocabulary threshold of 50.

NMT Systems We use the default configuration of the Transformer (Vaswani et al., 2017) NMT model implementation of the Sockeye NMT toolkit (Hieber et al., 2020). The exception is the use of source-side factors (Sennrich and Haddow, 2016) with the dimensionality of 8 for systems using TGA, which changes the model's combined source embedding dimensionality from 512 to 520. We train all models using early stopping with patience of 10 based on their development set perplexity (Prechelt, 1998).

Morphological Taggers The preparation of training data with TGA and WinoMT evaluation relies on the outputs of a morphological tagger. If the tagger produces biased outputs, the TGA annotations might become too noisy to be useful. Furthermore, a biased morphological tagger

Tagger	F1 masc.	F1 fem.
Paikens et al. (2013)	98.6	98.7
Stanza	94.7	95.1
UDPipe	92.5	92.4

Table 2: Performance of morphological taggers on gender feature classification evaluated on the Universal Dependencies test set.

could also render WinoMT evaluation unreliable. Thus we first benchmark several morphological taggers on grammatical gender feature classification. We use Latvian as a development language because of the availability of lexicon-based and data-driven morphological analysis tools. Specifically, we use the Universal Dependencies [1] test set to compare two data-driven tools – the Stanza toolkit (Qi et al., 2020) and UDPipe (Straka and Straková, 2017). Additionally, we evaluate a dictionary-based morphological analyser and statistical tagger[2] by Paikens et al. (2013). Table 2 gives F-1 scores on masculine and feminine feature tagging. Results indicate that none of the taggers exhibits salient bias in their tagging performance. As the only non-neural system yields better F-1 scores than the other two systems, we further compare Stanza and the tagger by Paikens et al. (2013) in their impact on BLEU and WinoMT metrics. Results indicated that the choice of the tagger does not have a notable effect on BLEU scores. In terms of WinoMT accuracy scores, the NMT system that was trained using TGA prepared with Stanza yields an accuracy that is about 3% better than the system using the tagger by Paikens et al. (2013). Thus, in all remaining experiments, we use the Stanza tagger as it provides pre-trained models for a wide range of languages.

TGA in Training Data   Preparing training data with TGA requires statistical word alignments between words of source and target language sentences and a target language morphological tagger. To obtain word alignments, we use *fast_align* (Dyer et al., 2013). To obtain grammatical gender information of target language words, we use the Stanza morphological tagger. When training NMT systems with TGA, we combine two copies of the original training data: one where all source-side

factors are set to `U` and the other containing TGA.

TGA During Inference   In training data, TGA annotate regular source language words with the grammatical gender information of corresponding target language words. We do not have access to the target language sentence during inference. Thus, we use co-reference resolution tools and extract the referential gender information from the source sentence instead. To do so, we first use co-reference resolution tools to obtain the co-reference graph. We then identify sub-graphs which contain gendered pronouns. Finally, we propagate the gender information within the graph and annotate the antecedents (see Figure 2). We set the annotations for the remaining unannotated words to `U`.

We use neural co-reference resolution tools by AllenNLP [3] (Lee et al., 2017) and Hugging Face[4] (based on work by Clark and Manning (2016)). We refer to these systems as **TGA AllenNLP** and **TGA HuggingFace** respectively. We also report the performance of NMT with TGA, when TGA use oracle information directly taken from WinoMT datasets and refer to these as **TGA Oracle**.

Evaluation   We evaluate general translation quality using the BLEU (Papineni et al., 2002) metric evaluated over WMT test sets. To calculate BLEU, we use SacreBLEU[5] (Post, 2018) on cased, detokenized data. Reference test sets are only pre-processed using Moses punctuation normalization script[6]. We use the WinoMT test suite (Stanovsky et al., 2019) to measure gender bias of our NMT systems.

4   Results and Discussion

Results from experiments evaluating gender bias using the WinoMT test suite are provided in Table 3. First, we observe that all baseline systems show a strong bias towards generating translations using masculine forms. The EN-RU baseline system is the most biased as it produces only one translation hypothesis with a feminine antecedent for every 8.4 hypotheses containing masculine antecedents. Meanwhile the EN-DE baseline system

---

[1]https://github.com/
UniversalDependencies/UD_Latvian-LVTB
[2]https://github.com/PeterisP/LVTagger

[3]https://github.com/allenai/allennlp
[4]https://github.com/huggingface/
neuralcoref
[5]SacreBLEU hash: `BLEU+case.mixed+numrefs.`
`1+smooth.exp+tok.13a+version.1.3.6`
[6]https://github.com/moses-smt/
mosesdecoder/blob/master/scripts/
tokenizer/normalize-punctuation.perl

WMT Data Systems				
	Acc.	**Δ G**	**Δ S**	**M:F**
EN-DE Baseline	66.7	10.2	14.4	2.6
TGA Oracle	89.0	-4.7	1.7	1
TGA HuggingFace	77.6	-0.1	11.9	1.6
TGA AllenNLP	81.5	-2.0	11.1	1.4
EN-FR Baseline	48.6	29.8	11.8	5.5
TGA Oracle	81.5	1.4	2.8	1.2
TGA HuggingFace	67.8	4.9	12.4	2
TGA AllenNLP	74.4	1.6	10.1	1.6
EN-LV Baseline	27.9	26.0	9.6	3.9
TGA Oracle	42.7	15.9	10.3	2.9
TGA HuggingFace	38.6	19.7	18.1	3.0
TGA AllenNLP	39.3	18.1	18.6	2.8
EN-LT Baseline	38.0	32.6	6.5	5.9
TGA Oracle	52.8	15.2	4.0	2.7
TGA HuggingFace	43.4	22.5	7.6	3.9
TGA AllenNLP	47.2	17.7	5.1	3.1
EN-RU Baseline	32.3	37.7	14.1	8.4
TGA Oracle	55.9	10.6	14.0	2.5
TGA HuggingFace	45.4	24.8	13.7	4.4
TGA AllenNLP	51.4	17.0	15.2	3.2
Proprietary Large Data System				
	Acc.	**Δ G**	**Δ S**	**M:F**
EN-LV Baseline	42.0	27.9	16.6	4.9
TGA Oracle	55.1	4.8	18.2	1.7
TGA HuggingFace	46.2	13.5	24.1	2.6
TGA AllenNLP	49.9	10.8	23.1	2.3

Table 3: Results on WinoMT test suite.

is the least biased with the M:F ratio being much lower – 2.6 (see the last column of Table 3). Our baseline systems for EN-DE, EN-FR and EN-RU language pairs, however, show comparable ΔG and WinoMT accuracy results to those reported by Stanovsky et al. (2019) for several publicly available commercial systems. These results confirm that our baselines, although being strongly biased, are not unordinary.

Results from experiments using TGA with oracle gender information show an improvement in WinoMT accuracy and ΔG for all language pairs (see Table 3 TGA Oracle). These results demonstrate that when training MT systems to use TGA reduces their reliance on gender stereotypes when information about the subject's gender is available, proving the usefulness of methods proposed here. Despite the availability of oracle gender information, none of the systems is entirely bias-free or obtains 100% accuracy. Thus methods proposed here could be combined with others, such as those proposed by Saunders and Byrne (2020), to achieve further improvements.

Effect on BLEU As expected, using TGA with reference sentence grammatical gender annotations has a positive effect on BLEU, thus confirming our hypothesis why and how the NMT system learns to rely on TGA as an additional source of information during training (see Table 4). It is equally important, however, that, when training NMT systems to use TGA, it does not degrade their performance when gender information is not necessary or is unavailable. Thus we test our systems for such cases by setting all TGA values to U and compare them to the baseline systems (see Table 4). To test for statistically significant differences between the results of NMT systems we use pairwise bootstrap resampling (Koehn, 2004) and significance threshold of 0.05. Results indicate no statistically significant differences between systems using uninformative TGA values and their baseline counterparts with an exception of results for EN-RU systems ($\Delta 0.4$ BLEU), which we find to be statistically significant.

Effect of Data Size To analyze gender bias and TGA performance depending on the quality and size of the training data, we use much larger EN-LV proprietary data (see Table 1) to train production-grade NMT systems and contrast them with EN-LV WMT data systems (see the two EN-LV sections in Table 3 and Table 5). First of all, we notice that although the large data baseline has higher WinoMT accuracy than the WMT data system, it has a similar ΔG. Decomposing ΔG as male and female grammatical gender F-1 scores (Table 5), however, clarifies that, although similarly skewed, the large data baseline has higher F-1 scores than the WMT data baseline. Next, we note, that larger training data size has a positive effect on the system's ability to use TGA more effectively as the large data system using TGA has a greater improvement on the two metrics measuring bias – ΔG and M:F[7] than its WMT data counterpart relative to its baseline. These findings suggest that TGA is a method that is applicable not only in small data settings but also in large data settings, such as commercial systems, for which it is even more effective.

Plugging-in Co-reference Resolution Tools Finally, we experiment with TGA using gender information provided by two off-the-shelf co-reference resolution tools, AllenNLP and Hugging Face. Re-

---

[7] ΔS results are not reliable or comparable when M:F ratios are large or differ by a large value. See result section of Saunders and Byrne (2020) for more discussion.

	Basline	TGA	All TGA= ʊ
EN-DE	45.4	49.5	45.3
EN-FR	36.6	40.9	36.4
EN-LV	16.6	18.9	17.0
EN-LT	14.8	16.6	14.7
EN-RU	27.1	31.6	26.7

Table 4: Comparison of test set performance measured in BLEU for Baseline systems and systems trained using TGA. TGA: performance when using reference sentence *grammatical gender* annotations. All TGA=U: performance when all annotations set to be unknown.

	Male			**Female**		
	F-1	P	R	F-1	P	R
WMT Data System						
Baseline	47.2	48.3	46.2	21.2	53.9	13.2
TGA Oracle	58.5	56.0	61.3	42.5	74.7	29.7
Proprietary Large Data System						
Baseline	58.8	50.8	69.7	30.9	70.3	19.8
TGA Oracle	66.9	65.8	68.0	62.1	83.3	49.5

Table 5: Results of antecedent translation. Reporting grammatical gender F-1 score, precision (P) and recall (R) for EN-LV systems trained on WMT and proprietary large data.

sults show that using TGA with either of the tools outperforms baseline systems for all languages pairs. Furthermore, TGA with gender information provided by AllenNLP shows only a 4.5 to 7.1% drop in WinoMT accuracy compared to results when using TGA with oracle information. To put this in perspective, Saunders and Byrne (2020) required a handcrafted gender-balanced profession set and additional rescoring models, for their EN-DE system to obtain comparable WinoMT accuracy and ΔG without loss of translation quality. In contrast, the methods proposed here require tools that are readily available, making them easily applicable in practice.

5 Conclusions

We proposed a method for training MT systems to use word-level annotations containing information about the subject's gender. To prepare training data, the method requires a morphological tagger to annotate regular source language words with grammatical gender information of the corresponding target language words. During inference, annotations can be used to provide information about

subjects' referential or social gender obtained by analyzing text beyond sentence boundaries or externally. In experiments with five language pairs, we showed that using such gender annotations reduces NMT systems' reliance on gender stereotypes in principle. We then further showed one way for how these findings can be used in practice by using off-the-shelf co-reference resolution tools.

The method proposed here decouples the task of acquiring the necessary gender information from the task of learning to translate correctly when such information is available. Thus system's ability to use such information can be achieved independently from its availability at training time. This allows for application-specific sources of gender information. Examples are the translation of chat or social media content, where users may choose to indicate their gender or translation of whole documents, where gender information may be obtained using annotations and anaphora resolution. Thus, we believe that the methods proposed here, will provide means to limit the propagation of gender stereotypes by NMT systems when translating into languages with grammatical gender.

The source code to reproduce our results for the publicly available data sets is published on GitHub[8].

References

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James R. Glass. 2018. Identifying and controlling important neurons in neural machine translation. *CoRR*, abs/1811.01157.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

[8] https://github.com/artursstaf/mitigating-gender-bias-wmt-2020

Judith Butler. 1990. Feminism and the subversion of identity. *New York and London: Routledge*, 3:1–25.

Kevin Clark and Christopher D Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262.

Jennifer Coates. 1987. *Women, men and language: A sociolinguistic account of gender differences in language*. Longman.

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35.

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

Mostafa Elaraby, Ahmed Y. Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender aware spoken language translation applied to english-arabic. *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6.

Pedro A Fuertes-Olivera. 2007. A corpus-based view of lexical gender in written business english. *English for Specific Purposes*, 26(2):219–234.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Marlis Hellinger and Heiko Motschenbacher. 2015. *Gender across languages*, volume 4. John Benjamins Publishing Company.

Felix Hieber, Tobias Domhan, Michael Denkowski, and David Vilar. 2020. Sockeye 2: A toolkit for neural machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 457–458, Lisboa, Portugal. European Association for Machine Translation.

Graeme Hirst. 1981. Anaphora in natural language understanding: A survey.

AI HLEG. 2019. Ethics guidelines for trustworthy ai. *High-Level Expert Group on Artificial Intelligence, Brussels*.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)*, pages 483–488.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the workshop on noisy user-generated text*, pages 9–18.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. Association for Computational Linguistics.

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54, Florence, Italy. Association for Computational Linguistics.

Peteris Paikens, Laura Rituma, and Lauma Pretkalniņa. 2013. Morphological analysis with limited resources: Latvian example. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 267–277.

636

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

Mārcis Pinnis, Andrejs Vasiļjevs, Rihards Kalniņš, Roberts Rozis, Raivis Skadiņš, and Valters Šics. 2018. Tilde MT platform for developing client specific MT solutions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2019. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, pages 1–19.

Lutz Prechelt. 1998. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

John R Rickford. 2016. *Raciolinguistics: How language shapes our ideas about race*. Oxford University Press.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. Association for Computational Linguistics.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13230–13241. Curran Associates, Inc.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.