First the worst: Finding better gender translations during beam search

Anonymous ACL submission

Abstract

Generating machine translations via beam search seeks the most likely output under a model. However, beam search has been shown to amplify demographic biases exhibited by a model. We aim to address this, focusing on gender bias resulting from systematic errors in grammatical gender translation. Almost all prior work on this problem adjusts the training data or the model itself. By contrast, our approach changes only the inference procedure.

We explore two techniques: applying constraints during inference to improve gender diversity in n-best lists, and reranking n-best lists using gender features obtained from the source sentence. Combining these methods gives large gains in gender translation accuracy for three language pairs without requiring additional bilingual data or retraining.

1 Introduction

Neural language generation models optimized by likelihood have a tendency towards 'safe' word choice. This lack of output diversity has been noted in NMT (Vanmassenhove et al., 2019) and throughout NLP (Li et al., 2016; Sultan et al., 2020). Model-generated language may be repetitive or stilted. More insidiously, generating the most likely output based only on corpus statistics can amplify any existing biases in the corpus (Zhao et al., 2017).

Potential harms arise when biases around word choice or grammatical gender inflections reflect demographic or social biases (Sun et al., 2019). The resulting gender mistranslations could involve implicit misgendering of a user or other referent, or perpetuation of social stereotypes about the 'typical' gender of a referent in a given context.

Dominant approaches to the problem almost exclusively involve retraining (Vanmassenhove et al., 2018; Escudé Font and Costa-jussà, 2019; Stafanovičs et al., 2020) or tuning (Saunders and Byrne, 2020; Basta et al., 2020) on gender-adjusted data. Such approaches are often computationally expensive and risk introducing new biases (Shah et al., 2020). Instead, this paper seeks to improve translations from existing models. Roberts et al. (2020) have recently highlighted beam search's tendency to amplify gender bias – we aim to guide it instead towards finding better gender translations.

Our contributions are as follows: we rerank the n-best lists of NMT models exhibiting gender bias, demonstrating that we can extract better gender translations from the *original model's* beam. We also generate new n-best lists subject to gendered inflection constraints, and show this increases the frequency of correctly gendered entities appearing in n-best lists. We make no changes to the NMT model or training data, and require only monolingual resources for the source and target languages.

1.1 Related work

Prior work mitigating gender bias in NLP often involves adjusting training data, directly (Zhao et al., 2018) or via embeddings (Bolukbasi et al., 2016). Our inference-only approach is closer to work on controlling or 'correcting' gendered output.

Controlling gender translation generally involves introducing external information into the model. Miculicich Werlen and Popescu-Belis (2017) integrate cross-sentence coreference links into reranking to improve pronoun translation. Vanmassenhove et al. (2018) and Moryossef et al. (2019) incorporate sentence-level gender features into training data and during inference respectively. Token-level source gender tags are used by Stafanovičs et al. (2020) and Saunders et al. (2020). As in this prior work, our focus is applying linguistic gender-consistency information, rather than obtaining it.

A separate line of work treats gender-related inconsistencies as a search and correction problem. Roberts et al. (2020) find that beam search amplifies gender bias compared to sampling search. Saunders and Byrne (2020) rescore translations with a model fine-tuned for additional gender sensitivity, constraining outputs to gendered-reinflections of the original. Related approaches for monolingual tasks reinflect whole-sentence gender (Habash et al., 2019; Alhafni et al., 2020; Sun et al., 2021). An important difference in our work is use of the same model for initial translation and reinflection, reducing computation and complexity.

2 Finding consistent gender in the beam

There are two elements to our proposed approach. First, we *produce an n-best list* of translations using our single model per language pair. We use either standard beam search or a two-pass approach where the second pass searches for differently-gendered versions of the highest likelihood initial translation. We then *select a translation* from the list, either by log likelihood or by how far the target language gender features correspond to the source sentence.

2.1 Gender-constrained n-best lists

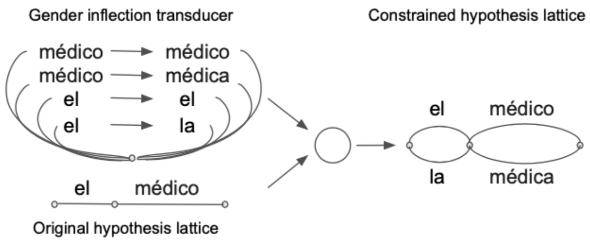

Figure 1: Constraints for a toy initial hypothesis.

We produce n-best lists in two ways. One option is standard beam search. Alternatively, we synthesize n-best lists using the gendered constraint scheme of Saunders and Byrne (2020), illustrated in Figure 1. This involves a second *gender-constrained* beam search pass to reinflect an initial hypothesis, producing a synthesized n-best list containing gendered alternatives of that hypothesis.

The second reinflection pass uses a target language *gender inflection transducer* which defines grammatically gendered reinflections. For example, Spanish definite article *el* could be unchanged or reinflected to *la*, and profession noun *médico* could be reinflected to *médica* (and vice versa). Composing the reinflections with the original hypothesis generates a *constrained hypothesis lattice*.

We can now perform constrained beam search, which can encourage NMT to output specific vocabulary (Stahlberg et al., 2016; Khayrallah et al., 2017). The only difference from standard beam search is that gender-constrained search only expands translations forming paths in the constrained hypothesis lattice. In the Figure 1 example, beam-n search would produce the n most likely translations, while the gender-constrained pass would only produce the 4 translations in the lattice.

Importantly, for each language pair we use just one NMT model to produce gendered variations of its *own* hypotheses. Unlike Saunders and Byrne (2020) we do not reinflect translations with a separate gender-sensitive model. This removes the complexity, potential bias amplification and computational load of developing the gender-translation-specific models central to their approach.

While we perform two full inference passes to simplify implementation, further efficiency improvements are possible. For example, the source sentence encoding could be reused for the reinflection pass. In principle, some beam search constraints could be applied in the first inference pass, negating the need for two passes. These potential efficiency gains would not be possible if using a separate NMT model to reinflect the translations.

2.2 Reranking gendered translations

Algorithm 1 Gender-reranking an n-best list

Input: x: Source sentence; Y: set of translation hypotheses for x; L: Log likelihoods for all $y \in Y$; A: word alignments between x and all y

$p, p_g \leftarrow$ pronoun_and_gender(x) ▷ Or oracle
$e \leftarrow$ get_entity(x, p) ▷ Or oracle
for all $y \in Y$ **do**
  $y_{score} \leftarrow 0$
  for all $t \in A_y(e)$ **do** ▷ Translated entity
    $t_g \leftarrow$ get_gender(t)
    if $t_g = p_g$ **then**
      y_{score} += 1
    end if
  end for
end for
$\hat{Y} = \{\text{argmax}_y(y_{score}, y \in Y)\}$
$\hat{y} = \text{argmax}_y(L(y), y \in \hat{Y})$
return $\hat{y}$

We select an output translation from an n-best list in two ways, regardless of whether the list was produced by beam search or the two-pass approach. One option selects the highest-likelihood translation under the NMT model. Alternatively, we rerank for gender consistency with the source sentence. We focus on either *oracle* or *inferred* entities coreferent with a source pronoun.

The *oracle* case occurs in several scenarios. Oracle entity labels could be provided as for the WinoMT challenge set (Stanovsky et al., 2019). They could also be user-defined for known entities (Vanmassenhove et al., 2018), or if translating the same sentence with different entity genders to produce multiple outputs (Moryossef et al., 2019).

The *inferred* case determines entities automatically. In Algorithm 1, pronoun_and_gender finds a source pronoun and its grammatical gender. We then find coreferent entities using a target language coreference resolution tool in get_entity. For brevity Algorithm 1 is written for one entity per sentence: in practice there is no such limit.

For each entity we find the aligned translated entity, similar to Stafanovičs et al. (2020). We determine the translated entity's grammatical gender by target language morphological analysis in get_gender. Finally we rerank, first by source gender agreement, tie-breaking with log likelihood[1].

3 Experimental setup

We translate English into German, Spanish and Hebrew using Transformers (Vaswani et al., 2017) with 30K BPE vocabularies (Sennrich et al., 2016). We train the en-de model on WMT19 newstask data including filtered Paracrawl (Barrault et al., 2019), en-es on UNCorpus data (Ziemski et al., 2016), and en-he on the IWSLT corpus (Cettolo et al., 2014). For further training details see Appendix A.

Some proposed steps require tools or resources: 1) For gender-constrained search, creating gender inflection transducers; 2) For inferred-reranking, finding source gendered entities 3) For all reranking, finding translated gendered entities; 4) For all reranking, getting translated entity genders.

For 1) we use Spacy (Honnibal and Montani, 2017) and DEMorphy (Altinok, 2018) morphological analysis for Spanish and German, and fixed rules for Hebrew, on large vocabulary lists to produce gender transducers, directly following Saunders and Byrne (2020)[2]. The highest likelihood outputs from beam-4 search form the original hypothesis lattices. For 2) we use a RoBERTa model (Liu et al., 2019) tuned for coreference on Winograd challenge data[3]. For 3) we use fast_align (Dyer et al., 2013). For 4) we use the same morphological analysis as in 1, now on translated entities.

We evaluate gender translation on WinoMT (Stanovsky et al., 2019), measuring overall accuracy and ΔG (F1 score difference between masculine and feminine labelled sentences, closer to 0 is better). As WinoMT lacks references we assess cased BLEU on WMT18 (en-de, 3K sentences), WMT13 (en-es, 3K sentences) and IWSLT14 (en-he, 962 sentences) using SacreBLEU (Post, 2018). For validation during NMT model training we use test sets from earlier years of the same tasks.

4 Results and discussion

We first discuss the possibilities of oracle-reranking n-best lists in Table 1, before proceeding to the more general scenario of inferred-reranking. Comparing lines 1 vs 2, gender-constrained beam-4 search scores very closely to standard beam-4 search for all metrics and language pairs if simply taking the highest likelihood output. For beam-20 (5 vs 6) en-de and en-es, constraints do mitigate the BLEU degradation common with larger beams (Stahlberg and Byrne, 2019).

In lines 1 vs 3, 5 vs 7, we oracle-rerank beam search outputs instead of choosing by highest likelihood. We see about 10% accuracy improvement relative to non-reranked beam-4 across languages, and over 25% relative improvement for beam-20. Combining oracle-reranking and constraints further boosts accuracy. This suggests constraints encourage presence of better gender translations in n-best lists, but that reranking is needed to extract them.

Using beam-20 significantly improves the performance of reranking. With constraints, beam-20 oracle-reranking gives *absolute* accuracy gains of about 20% over the highest likelihood beam search output. However, beam-4 shows most of the improvement over that baseline. We find diminishing returns as beam size increases (Appendix B), suggesting large, expensive beams are not necessary.

So far we have shown accuracy improvements with oracle reranking, indicating that the synthesized n-best lists often contain a gender-accurate hypothesis. In Table 2, we explore inferred-reranking using a RoBERTa model, investigating whether that hypothesis can be found automatically. We find very little degradation in WinoMT accu-

[1]Reranking code and n-best lists included with submission.

[2]Scripts and data for lattice construction as in Saunders and Byrne (2020) were provided by those authors at `https://github.com/DCSaunders/gender-debias`

[3]Model from `https://github.com/pytorch/fairseq/tree/master/examples/roberta/wsc`

	Beam	Gender constrain	Oracle rerank	en-de			en-es			en-he		
				BLEU	Acc	ΔG	BLEU	Acc	ΔG	BLEU	Acc	ΔG
1	4	×	×	**42.7**	60.1	18.6	27.5	47.8	38.4	23.8	47.5	21.1
2		✓	×	**42.7**	59.1	20.1	**27.8**	48.3	36.2	23.8	47.4	21.5
3		×	✓	-	66.5	10.1	-	53.9	25.9	-	52.0	16.8
4		✓	✓	-	77.9	**-0.6**	-	55.7	22.3	-	54.5	13.7
5	20	×	×	42.3	59.0	20.1	27.3	46.4	40.7	**24.0**	46.8	22.5
6		✓	×	**42.7**	59.0	20.3	**27.8**	48.3	36.2	23.8	47.3	21.7
7		×	✓	-	74.3	2.4	-	63.5	11.0	-	59.3	11.2
8		✓	✓	-	**84.2**	-3.6	-	**66.3**	**8.1**	-	**65.3**	**4.9**

Table 1: Accuracy (%) and masculine/feminine F1 difference ΔG, oracle-reranking WinoMT. BLEU scores are for en-de WMT18, en-es WMT13, and en-he IWSLT14, which lack gender labels so cannot be oracle-reranked.

	Beam	Gender constrain	Inferred rerank	en-de			en-es			en-he		
				BLEU	Acc	ΔG	BLEU	Acc	ΔG	BLEU	Acc	ΔG
1	4	×	✓	42.7	65.9	10.7	27.5	52.6	28.1	23.8	51.3	17.0
2		✓	✓	42.7	76.4	0.5	27.8	53.9	24.6	23.8	53.6	14.4
3	20	×	✓	42.2	72.9	3.3	27.3	60.2	15.3	24.0	57.8	11.9
4		✓	✓	42.6	81.8	-2.6	27.8	63.5	10.9	23.8	62.8	6.2

Table 2: Accuracy (%) and masculine/feminine F1 difference ΔG. Inferred-reranking with genders and entities for WinoMT and generic test sets determined by a RoBERTa model. Non-reranked results unchanged from Table 1.

Beam	System	en-de	en-es	en-he
4	S&B	79.4	62.2	53.1
	S&B + rerank	81.9	68.9	56.6
20	S&B	79.6	62.1	52.8
	S&B + rerank	83.6	73.9	62.9

Table 3: WinoMT accuracy for inferred-reranking the adaptation scheme of Saunders and Byrne (2020).

racy when inferring entities compared to the oracle (Table 1). We hypothesise that difficult sentences are hard for both coreference resolution and NMT, so cases where RoBERTa disambiguates wrongly are also mistranslated with oracle information.

We are unable to oracle-rerank the generic test sets, since they have no oracle gender labels. However, we can tag them using RoBERTA for inferred-reranking. In Table 2 we find this has little or no impact on BLEU score, unsurprising for sets not designed to highlight potentially subtle gender translation effects. This suggests positively that our scheme does not impact general translation quality.

So far we have not changed the NMT model at all. In Table 3, for comparison, we investigate the approach of Saunders and Byrne (2020): tuning a model on their dataset of gendered profession sentences, then constrained-rescoring the original model's hypotheses.[4] We do indeed see strong gender accuracy improvements with their approach, but inferred-reranking the resulting models' n-best lists further improves scores. We also note that line 4 of Table 2 has higher accuracy than non-reranked beam-20 S&B, without any model fine-tuning.

Finally, we note that our approach can be used for disambiguation in many gender translation scenarios. One benefit beyond the scope of this paper is application to incorporating context: gender information could potentially come from a prior or subsequent sentence, or reflect external user preferences. Another benefit is flexibility to introducing new gendered vocabulary such as neopronouns: if a term can be defined for reranking, it can be searched for in the beam without retraining.

5 Conclusions

This paper attempts to improve gender translation without a single change to the NMT model. We demonstrate that gender-constraining the target language during inference can encourage models to produce n-best lists with correct hypotheses. Moreover, we show that simple reranking heuristics can extract more accurate gender translations from the n-best lists using oracle or inferred information.

Unlike other approaches to this problem we do not attempt to counter unidentified and potentially intractable sources of bias in the training data, or produce new models. However, our approach does significantly boost the accuracy of a prior data-centric bias mitigation technique. In general we view our scheme as orthogonal to such approaches: if a model ranks diverse gender translations higher in the beam initially, finding better gender translations during beam search becomes simpler.

[4] Different scores from the original authors may be due to variations in hyperparameters, or WinoMT updates.

Impact statement[5]

Where machine translation is used in people's lives, mistranslations have the potential to misrepresent people. This is the case when personal characteristics like social gender conflict with model biases towards certain forms of grammatical gender. As mentioned in the introduction, the result can involve implicit misgendering of a user or other human referent, or perpetuation of social biases about gender roles as represented in the translation. A user whose words are translated with gender defaults that imply they hold such biased views will also be misrepresented.

We attempt to avoid these failure modes by identifying translations which are at least consistent within the translation and with the source sentence. This is dependent on identifying grammatically gendered terms in the target language – however, this element is very flexible and can be updated for new gendered terminology. We note that models which do not account for variety in gender expression such as neopronoun use may not be capable of generating appropriate gender translations, but in principle a variety of gender translations could be extracted from the beam.

By avoiding the data augmentation, tuning and retraining elements in previously proposed approaches to gender translation, we seek to simplify the process and remove additional stages during which bias could be introduced or amplified (Shah et al., 2020).

In terms of compute time and power, we minimize impact by using a single GPU only for training the initial NMT models exactly once for the iterations listed in Appendix A. All other experiments involve rescoring those models and run in parallel on CPUs in under an hour, except the experiments following Saunders and Byrne (2020), an approach itself involving only minutes of GPU fine-tuning.

[5]Impact statement permitted on 5th page by ARR guidelines

References

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2020. Gender-aware reinflection using linguistically enhanced neural models. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 139–150, Barcelona, Spain (Online). Association for Computational Linguistics.

Duygu Altinok. 2018. DEMorphy, German language morphological analyzer. *arXiv preprint arXiv:1803.00902*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 99–102, Seattle, USA. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign, IWSLT 2014. In *Proceedings of the International Workshop on Spoken Language Translation, Hanoi, Vietnam*, page 57.

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, Atlanta, Georgia. Association for Computational Linguistics.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

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.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with bloom embeddings. *Convolutional Neural Networks and Incremental Parsing*.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. In *Proceedings of the International Workshop on Spoken Language Translation 2016*, volume 1. 13th International Workshop on Spoken Language Translation 2016, IWSLT 2016.

Huda Khayrallah, Gaurav Kumar, Kevin Duh, Matt Post, and Philipp Koehn. 2017. Neural lattice search for domain adaptation in machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 20–25, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Using coreference links to improve Spanish-to-English machine translation. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 30–40, Valencia, Spain. 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.

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, Belgium, Brussels. Association for Computational Linguistics.

Nicholas Roberts, Davis Liang, Graham Neubig, and Zachary C Lipton. 2020. Decoding and diversity in machine translation. *arXiv preprint arXiv:2011.13477*.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). 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.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation (WMT)*.

Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3356–3362, Hong Kong, China. Association for Computational Linguistics.

Felix Stahlberg, Eva Hasler, Aurelien Waite, and Bill Byrne. 2016. Syntactically guided neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 299–305, 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.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656, Online. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english. *arXiv preprint arXiv:2102.06788*.

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, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 193–199, Boston, MA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Model training details

All NMT models are 6-layer Transformers with a 30K BPE vocabulary, trained using Tensor2Tensor with batch size 4K (Vaswani et al., 2018). All data except Hebrew is truecased and tokenized using (Koehn et al., 2007). The en-de model is trained for 300K batches, en-es for 150K batches, and en-he for 15K batches, transfer learning from the en-de model. We filter the BPE-ized data for maximum (80) and minimum (3) length, and length ratio 3.

B Beam size for constrained reranking

In this paper we present results with beam sizes 4 and 20. Beam-4 search is commonly-used and meets a speed-quality trade-off for NMT (see e.g. Junczys-Dowmunt et al. (2016)). Beam-20 is still practical, but approaches diminishing returns for improved quality without search error mitigation (Stahlberg and Byrne, 2019). These beam sizes therefore illustrate contrasting levels of practical reranking. However, it is instructive to consider how wide a beam is really needed for improvements under gender-constrained reranking.

In Figure 2 we report WinoMT accuracy under gender-constrained oracle reranking with beam width increasing by intervals of 4. For all systems, the largest jump in improvement is between beam sizes 4 and 8, with diminishing returns after beam-12. The en-de curve is relatively shallow, possibly due to strong scores before reranking, or even a performance ceiling determined by the WinoMT framework itself. Curves for en-he and en-es are very close, suggesting a similarity between the gender distribution in the n-best lists for those models.

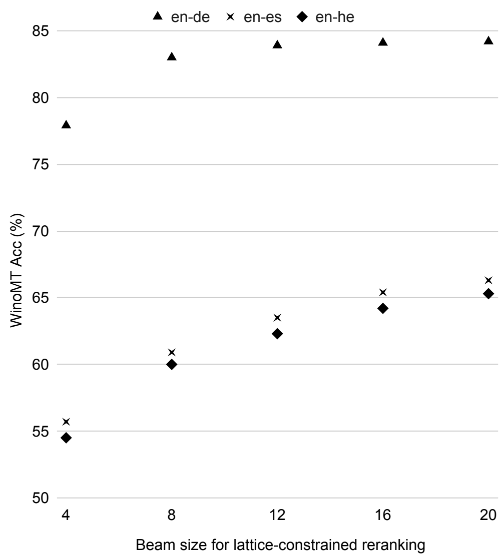

Figure 2: WinoMT accuracy after oracle-reranking gender-constrained n-best lists, varying n.

C Constrained vs unconstrained beams

We can observe the difference between standard and constrained beam search by examining the n-best lists. Table 4 (next page) gives 5 examples of 4-best lists for WinoMT sentences translated into German. Examples are not cherry-picked but selected from throughout WinoMT with a random number generator. Lists are ordered by NMT model likelihood and produced with standard unconstrained beam search, and with constrained beam search.

With standard unconstrained beam search, translations vary primarily in words unrelated to the entities, such as synonyms or verb tenses. However, entity grammatical genders are generally unchanged throughout the unconstrained n-best lists, except for 1 where the secondary entity changes. Reranking cannot always find a gender-consistent translation in the unconstrained lists, defaulting to the 1best for all except 2 (which seems to have a poorly aligned hypothesis).

By contrast, constrained beam search ensures the n-best list contains gendered inflections of the initial best-scoring translation. The changes vary the grammatical genders of articles and entities, resulting in more gender-diverse hypotheses, and allowing reranking to find a better translation for 1.

We note that in 3, 4 and 5 both the pronoun and the default gender convention for unknown gender entities are masculine. Reranking is not strictly necessary to find a better translation for these sentences, since the highest likelihood output is gender-consistent. However, we note that some outputs with gender constraints do in fact vary the gender of the secondary entity – the entity with unspecified gender. This illustrates our approach's ability to improve n-best list diversity even when it does not necessarily impact translation consistency.

We observe occasional grammatical inconsistencies in n-best hypotheses (e.g. "die Fahrer" in 3). When constraining beam search to grammatical variations of a sentence with an imperfect NMT model, we expect some hypotheses with grammatical degradation. However, our priority, and the purpose of our reranking scheme, is consistency with the source in the output translation, not inconsistencies elsewhere in the n-best list.

1. The broker called the client because she had new stocks to recommend.	
Unconstrained	* *Der Makler* rief *den Kunden* an, weil sie neue Aktien zu empfehlen hatte. *Der Makler* rief die Klientin an, weil sie neue Aktien zu empfehlen hatte. *Der Makler* rief *den Klienten* an, weil sie neue Aktien zu empfehlen hatte. *Der Makler* rief *den Kunden* an, weil *er* neue Aktien zu empfehlen hatte.
Constrained	*Der Makler* rief *den Kunden* an, weil sie neue Aktien zu empfehlen hatte. *Der Makler* rief *den Kunden* an, weil *er* neue Aktien zu empfehlen hatte. * Die Maklerin rief *den Kunden* an, weil sie neue Aktien zu empfehlen hatte. *Der Makler* rief die *Kunden* an, weil sie neue Aktien zu empfehlen hatte.
2. The teacher explained the situation to the chief and felt understood by her.	
Unconstrained	Die Lehrerin erklärte die Situation *dem Chef* und fühlte sich von ihr verstanden. * Die Lehrerin erklärte *dem Obersten* die Situation und fühlte sich von ihr verstanden. Die Lehrerin erklärte die Situation *dem Obersten* und fühlte sich von ihr verstanden. Die Lehrerin erläuterte die Situation *dem Obersten* und fühlte sich von ihr verstanden.
Constrained	*Die Lehrerin erklärte die Situation *dem Chef* und fühlte sich von ihr verstanden. *Der Lehrer* erklärte die Situation *dem Chef* und fühlte sich von ihr verstanden. *Der Lehrer* erklärte die Situation *dem Chef* und fühlte sich von *ihm* verstanden. Die Lehrerin erklärte die Situation *dem Chef* und fühlte sich von *ihm* verstanden.
3. The assistant hired *the driver* as *he* offered an excellent service.	
Unconstrained	* *Der Assistent* stellte *den Fahrer* ein, da *er* einen ausgezeichneten Service bot. *Der Assistent* stellte *den Fahrer* ein, da *er* einen exzellenten Service bot. *Der Assistent* stellte *den Fahrer* ein, da *er* einen hervorragenden Service bot. *Der Assistent* stellte *den Fahrer* ein, als *er* einen ausgezeichneten Service bot.
Constrained	* *Der Assistent* stellte *den Fahrer* ein, da *er* einen ausgezeichneten Service bot. Die Assistentin stellte *den Fahrer* ein, da *er* einen ausgezeichneten Service bot. *Der Assistent* stellte die *Fahrer* ein, da *er* einen ausgezeichneten Service bot. *Der Assistent* stellte *den Fahrer* ein, da *er* eine ausgezeichnete Service bot.
4. *The laborer* handed the application to the editor because *he* wants the job.	
Unconstrained	* *Der Arbeiter* übergab den Antrag an *den Herausgeber*, weil *er* den Job will. *Der Arbeiter* übergab die Bewerbung an *den Herausgeber*, weil *er* die Arbeit will. *Der Arbeiter* übergab die Bewerbung an *den Herausgeber*, weil *er* den Job will. *Der Arbeiter* überreichte die Bewerbung an *den Herausgeber*, weil *er* die Arbeit will.
Constrained	* *Der Arbeiter* übergab den Antrag an *den Herausgeber*, weil *er* den Job will. *Der Arbeiter* übergab den Antrag an *den Herausgeber*, weil *er* diesen Job will. *Der Arbeiter* übergab den Antrag an *den Herausgeber*, weil *er* die Job will. *Der Arbeiter* übergab die Antrag an *den Herausgeber*, weil *er* die Job will.
5. *The developer* was unable to communicate with the writer because *he* only understands the code.	
Unconstrained	* *Der Entwickler* konnte nicht mit *dem Schriftsteller* kommunizieren, weil *er* nur den Code versteht. *Der Entwickler* konnte nicht mit *dem Autor* kommunizieren, weil *er* nur den Code versteht. *Der Entwickler* war nicht in der Lage, mit *dem Schriftsteller* zu kommunizieren, weil *er* nur den Code versteht. *Der Entwickler* war nicht in der Lage, mit *dem Autor* zu kommunizieren, weil *er* nur den Code versteht.
Constrained	* *Der Entwickler* konnte nicht mit *dem Schriftsteller* kommunizieren, weil *er* nur den Code versteht. *Der Entwickler* konnte nicht mit der Schriftstellerin kommunizieren, weil *er* nur den Code versteht. *Der Entwickler* konnte nicht mit *dem Schriftsteller* kommunizieren, weil *er* nur die Code versteht. *Der Entwickler* konnte nicht mit *dem Schriftsteller* kommunizieren, weil *er* nur diesen Code versteht.

Table 4: English-German 4-best lists for 5 randomly-selected WinoMT sentences, translated with normal beam search and gender-constrained beam search. Grammatically feminine human entities are underlined. Grammatically masculine human entities are *emphasised*. Lists are ordered by NMT model likelihood (first is 1best) - lines marked with * are those selected under oracle-reranking.
1: Constrained reranking finds a better gender translation that is not present in the unconstrained beam.
2: A better gendered translation is not found in either width-4 beam. Constraints still maintain semantic meaning throughout the beam while allowing syntactic variation, including a differently gendered secondary entity.
3, 4, 5: The highest likelihood output is acceptable. For 3 and 5 constraining the n-best list results in more gender variation.