Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques

Joel Escudé Font and Marta R. Costa-jussà
Universitat Politècnica de Catalunya, 08034 Barcelona
joel.escude@estudiant.upc.edu,marta.ruiz@upc.edu

Abstract

Neural machine translation has significantly pushed forward the quality of the field. However, there are remaining big issues with the output translations and one of them is fairness. Neural models are trained on large text corpora which contain biases and stereotypes. As a consequence, models inherit these social biases. Recent methods have shown results in reducing gender bias in other natural language processing tools such as word embeddings. We take advantage of the fact that word embeddings are used in neural machine translation to propose a method to equalize gender biases in neural machine translation using these representations. We evaluate our proposed system on the WMT English-Spanish benchmark task, showing gains up to one BLEU point. As for the gender bias evaluation, we generate a test set of occupations and we show that our proposed system learns to equalize existing biases from the baseline system.

1 Introduction

Language is one of the most interesting and complex skills used in our daily life, and may even be taken for granted on our ability to communicate. However, the understanding of meanings between lines in natural languages is not straightforward for the logic rules of programming languages.

Natural language processing (NLP) is a subfield of artificial intelligence that focuses on making natural languages understandable to computers.

Similarly, the translation between different natural languages is a task for Machine Translation (MT). Neural MT has shown significant improvements on performance using deep learning techniques, which are algorithms that learn abstractions from data. In recent years, these deep learning techniques have shown promising results in narrowing the gap between human-like performance with sequence-to-sequence learning approaches in a variety of tasks (Sutskever et al., 2014), improvements in combination of approaches such as attention (Bahdanau et al., 2014) and translation systems algorithms like the Transformer (Vaswani et al., 2017).

One downside of models trained with human generated corpora is that social biases and stereotypes from the data are learned (Madaan et al., 2018). A systematic way of showing this bias is by means of word embeddings, a vector representation of words. The presence of biases, such as gender bias, is studied for these representations and evaluated on crowd-sourced tests (Bolukbasi et al., 2016). The presence of biases in the data can directly impact downstream applications (Zhao et al., 2018a) and are at risk of being amplified (Zhao et al., 2017).

The objective of this work is to study the presence of gender bias in MT and give insight on the impact of debiasing in such systems. An example of this gender bias is the word “friend” in the English sentence “She works in a hospital, my friend is a nurse” would be correctly translated to “amiga” (girl friend in Spanish) in Spanish, while “She works in a hospital, my friend is a doctor” would be incorrectly translated to “amigo” (boy friend in Spanish) in Spanish. We consider that this translation contains gender bias since it ignores the fact that, for both cases, “friend” is a female and translates by focusing on the occupational stereotypes, i.e. translating doctor as male and nurse as female.

The main contribution of this study is providing progress on the recent detected problem which is gender bias in MT (Prates et al., 2018). The progress towards reducing gender bias in MT is made in two directions: first, we define a frame-

Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing, pages 147–154
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

work to experiment, detect and evaluate gender bias in MT for a particular task; second, we propose to use debiased word embeddings techniques in the MT system to reduce the detected bias. This is the first study in proposing debiasing techniques for MT.

The rest the paper is organized as follows. Section 2 reports material relevant to the background of the study. Section 3 presents previous work on the bias problem. Section 4 reports the methodology used for experimentation and section 5 details the experimental framework. The results and discussion are included in section 6 and section 7 presents the main conclusions and ideas for further work.

2 Background

This section presents the models used in this paper. First, we describe the Transformer model which is the state-of-the-art model in MT. Second, we report describe word embeddings and, then, the corresponding techniques to debias them.

2.1 Transformer

The Transformer (Vaswani et al., 2017) is a deep learning architecture based on self-attention, which has shown better performance over previous systems. It is more efficient in using computational resources and has higher training speed than previous recurrent (Sutskever et al., 2014; Bahdanau et al., 2014) and convolutional models (Gehring et al., 2017).

The Transformer architecture consists of two main parts: an encoder and a decoder. The encoder reads an input sentence to generate a representation which is later used by a decoder to produce a sentence output word by word.

The input words are represented as vectors, word embeddings (more on this in section 2.2) and then, positional embeddings keep track of the sequentiality of language. The Transformer architecture computes a reduced constant number of steps using a self-attention mechanism on each one. The attention score is computed for all words in a sentence when comparing the contribution of each word to the next representation. New representations are generated in parallel for all words at each step .

Finally, the decoder uses self-attention in generated words and also uses the representations from the last words in the encoder to produce a single word each time.

2.2 Word embeddings

Word embeddings are vector representations of words. These representations are used in many NLP applications. Based on the hypothesis that words appearing in same contexts share semantic meaning, this continuous vector space representation gathers semantically similar words, thus being more expressive than other discrete representations like one-hot vectors.

Arithmetic operations can be performed with these embeddings, in order to find analogies between pairs of nouns with the pattern “A is to B what C is to D” (Mikolov et al., 2013). For nouns, such as countries and their respective capitals or for the conjugations of verbs.

While there are many techniques for extracting word embeddings, in this work we are using Global Vectors, or GloVe (Pennington et al., 2014). Glove is an unsupervised method for learning word embeddings. This count-based method, uses statistical information of word occurrences from a given corpus to train a vector space for which each vector is related to a word and their values describes their semantic relations.

2.3 Equalizing biases in word embeddings

The presence of biases in word embeddings is a topic of discussion about fairness in NLP. More specifically, Bolukbasi et al. (2016) proposes a post-process method for debiasing already trained word embeddings. (Zhao et al., 2018b) aims to restrict learning biases during the training of the embeddings to obtain a more neutral representation. The main ideas behind these methods are described next.

Hard-debiased embeddings (Bolukbasi et al., 2016) is a post-process method for debiasing word embeddings. First, the direction of the embeddings where the bias is present is identified. Second, the gender neutral words in this direction are neutralized to zero and also equalizes the sets by making the neutral word equidistant to the remaining ones in the set. The disadvantage of the first part of the process is that it can remove valuable information in the embeddings for semantic relations between words with several meanings that are not related to the bias being treated.

GN-GloVe (Zhao et al., 2018b) is an algorithm for learning gender neutral word embeddings models. It is based on the GloVe representation (Pennington et al., 2014) and modified to learn such word representations while restricting specific attributes, such as gender information, to specific dimensions. A set of seed male and female words are used to define metrics for computing the optimization and a set of gender neutral words is used for restricting neutral words in a gender direction.

3 Related work

While there are many studies on the presence of biases in many NLP applications, studies of this type in MT are quite limited.

Prates et al. (2018) performs a case study on gender bias in machine translation. They build a test set consisting of a list of jobs and gender-specific sentences. Using English as a target language and a variety of gender neutral languages as a source, i.e. languages that do not explicitly give gender information about the subject, they test these sentences on the translating service Google Translate. They find that occupations related to science, engineering and mathematics present a strong stereotype toward male subjects.

Vanmassenhove et al. (2018) compile a large multilingual dataset on the politics domain that contains the speaker information. They specifically use this information to incorporate it in a MT system. Adding this information improves the translation quality.

Our contribution is different from previous approaches in the sense that we are explicitly proposing a gender-debiased approach for NMT as well as an specific analysis based on correference and stereotypes to evaluate the effectiveness of our technique.

4 Methodology

In this section, we describe the methodology used for this study. The prior layer of both the encoder and decoder in the Transformer (Vaswani et al., 2017), where the word embeddings are trained, is adapted to use pre-trained word embeddings. We train the system with different pre-trained word embeddings (based on GloVe (Pennington et al., 2014)) to have a set of models. The scenarios are the following:

- No pre-trained word embeddings, i.e. they are learned within the training of the model.
- Pre-trained word embeddings learned from the same corpus. Specifically, GloVe, Hard-Debiased GloVe and Gender Neutral Glove (GN-GloVe) embeddings.

Also, the models with pre-trained embeddings given to the Transformer have three cases: using pre-trained embeddings only in the encoder side, see Figure 1 (left), only in the decoder side, Figure 1 (center), and both in the encoder and decoder sides, Figure 1 (right).

5 Experimental framework

In this section, we present the experimental framework. We report details on the training of the word embeddings and the translation system. We describe the data related to the training corpus and test sets and the parameters. Also, we comment on the use of computational resources.

5.1 Corpora

The language pair used for the experiments is English-Spanish. The training set consists of 16,554,790 sentences from a variety of sources including United Nations (Ziemski et al., 2016), Europarl (Koehn, 2005), CommonCrawl and News available from the Workshop on Machine Translation (WMT) [1]. The validation and test sets used are the *newstest2012* (3,003 sentences) and *newstest2013* (3,000 sentences), respectively, also from the same WMT workshop. See Table 2 for the corpus statistics.

To study gender bias, we have developed an additional test set with custom sentences to evaluate the quality of the translation in the models. We built this test set using a sentence pattern "*I've known* {*her, him, <proper noun>*} *for a long time, my friend works as* {*a, an*} *<occupation>*." for a list of occupations from different professional areas. We refer to this test as *Occupations test*, their related sizes are also listed in Table 2 and sample sentences from this set are in Table 1. We use Spanish proper names to reduce ambiguity in this particular test. These sentences are properly tokenized before using them in the test.

With these test sentences we see how "friend" is translated into its Spanish equivalent "amiga" or "amigo" which has a gender relation for each word, female and male, respectively. Note that we are formulating sentences with an ambiguous

[1] http://www.statmt.org/wmt13/

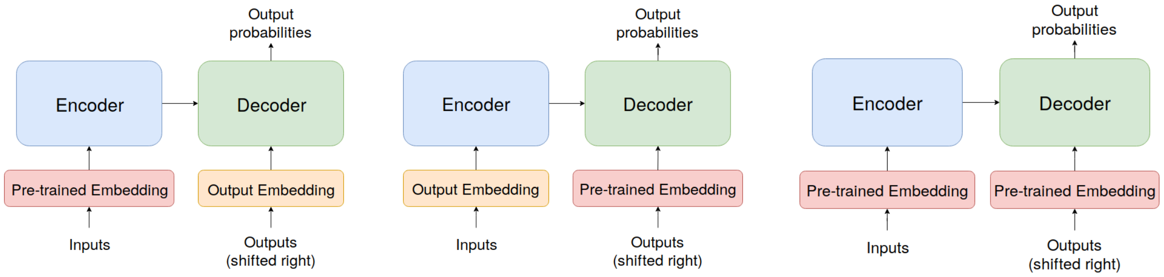

Figure 1: (Left) Pre-trained word embeddings in the encoder. (Center) Pre-trained word embeddings in the decoder. (Right) Pre-trained word embeddings in both the encoder and the decoder.

word "friend" that can be translated into any of the two words and we are adding context in the same sentence so that the system has enough information to translate them correctly. The list of occupations is from the U.S. Bureau of Labor Statistics[2], which also includes statistical data for gender and race for most professions. We use a pre-processed version of this list from (Prates et al., 2018).

5.2 Models

The architecture to train the models for the translation task is the Transformer (Vaswani et al., 2017) and we used the implementation provided by the OpenNMT toolkit[3]. The parameter values used in the Transformer are the same as proposed in the OpenNMT baseline system. Our baseline system is the Transformer witout pre-trained word embeddings.

Additionally, OpenNMT has built-in tools for training with pre-trained embeddings. In our case, these pre-trained embeddings have been implemented with the corresponding github repositories in GloVe[4], Hard-Debiasing with Debiaswe[5] and GN-GloVe[6].

The GloVe and GN-GloVe embeddings are trained from the same corpus presented in the previous section. We refer to the method from Bolukbasi et al. (2016) applied to the previously mentioned GloVe embeddings as Hard-Debiased GloVe. The dimension of the vectors is settled to 512 (as standard) and kept through all the experiments in this study. The parameter values for training the word embedding models are shown in Table 3.

Bolukbasi et al. (2016) uses a set of words to define the gender direction and to neutralize and equalize the bias from the word vectors. Three set of words are used: One set of ten pairs of words such as *woman-man, girl-boy, she-he* are used to define the gender direction. Another set of 218 gender-specific words such as *aunt, uncle, wife, husband* are used for learning a larger set of gender-specific words. Finally, a set of crowd-sourced male-female equalization pairs such as *dad-mom, boy-girl, granpa-grandma* that represent gender direction are equalized in the algorithm. In fact, for the English side, the gendered pairs used are the same as identified in the crowd-sourcing test by Bolukbasi et al. (2016). For the Spanish side, the sets are translated manually and modified when necessary to avoid non-applicable pairs or unnecessary repetitions. The sets from Zhao et al. (2018b) are similarly adapted to the Spanish language.

To evaluate the performance of the models we use the BLEU metric (Papineni et al., 2002). This metric gives a score for a predicted translation set compared to its expected output.

5.3 Hardware resources

The GPUs used for training are separate groups of four NVIDIA TITAN Xp and NVIDIA GeForce GTX TITAN. The duration time for training is approximately 3 and 5 days, respectively. In the implementation, the model is set to accumulate the gradient two times before updating the parameters, which simulates 4 more GPUs during training giving a total of 8 GPUs.

6 Results

In this section we report results on translation quality and present an analysis on gender bias.

[2] https://www.bls.gov/cps/tables.htm#empstat
[3] http://opennmt.net/
[4] https://github.com/stanfordnlp/GloVe
[5] https://github.com/tolga-b/debiaswe
[6] https://github.com/uclanlp/gn_glove

(En) I’ve known *her* for a long time, my *friend* works as an *accounting clerk*. (Es) *La* conozco desde hace mucho tiempo, mi *amiga* trabaja como *contable*.
(En) I’ve known *him* for a long time, my *friend* works as an *accounting clerk*. (Es) *Lo* conozco desde hace mucho tiempo, mi *amigo* trabaja como *contable*.
(En) I’ve known *Mary* for a long time, my *friend* works as an *accounting clerk*. (Es) Conozco a *Mary* desde hace mucho tiempo, mi *amiga* trabaja como *contable*.
(En) I’ve known *John* for a long time, my *friend* works as an *accounting clerk*. (Es) Conozco a *John* desde hace mucho tiempo, mi *amigo* trabaja como *contable*.

Table 1: Sample sentences from the *Occupations test* set. English (En) and Spanish (Es).

Language	Data set	Num. of sentences	Num. of words	Vocab. size
English (En)	Train	16.6M	427.6M	1.32M
	Dev	3k	73k	10k
	Test	3k	65k	9k
	Occupations test	1k	17k	0.8k
Spanish (Es)	Train	16.6M	477.3M	1.37M
	Dev	3k	79k	12k
	Test	3k	71k	11k
	Occupations test	1k	17k	0.8k

Table 2: English-Spanish data set.

Parameter	Value
Vector size	512
Memory	4.0
Vocab. min. count	5
Max. iter.	15
Window size	15
Num. threads	8
X max.	10
Binary	2
Verbose	2

Table 3: Word Embeddings Parameters.

Baseline			29.78
Pre-trained emb.	Enc.	Dec.	Enc./Dec.
GloVe	30.21	30.24	30.62
GloVe Hard-Deb.	30.16	30.09	29.95
GN-GloVe	29.12	30.13	**30.74**

Table 4: BLEU scores for the *newstest2013* test set. English-Spanish. Pre-trained embeddings are updated during training. In bold best results.

6.1 Translation

For the test set *newstest2013*, BLUE scores are given in Table 4. Pre-trained embeddings are used for training in three scenarios: in the encoder side (Enc.), in the decoder side (Dec.) and in both the encoder and decoder sides (Enc./Dec.). These pre-trained embeddings are updated during training. We are comparing several pre-trained embeddings against a baseline system (‘Baseline’ in Table 4) which does not include pre-trained embeddings (neither on the encoder nor the decoder).

For the studied cases, values do not differ much. Using pre-trained embeddings can improve the translation, which is coherent with previous studies (Qi et al., 2018). Furthermore, debiasing with GN-GloVe embeddings keeps this improvement and even increases it when used in both the encoder and decoder sides. We want to underline that these models do not decrease the quality of translation in terms of BLEU when tested in a standard MT task. Next, we show how each of the models performs on a gender debiasing task.

6.2 Gender Bias

A qualitative analysis is performed on the *Occupations test* set. Examples of this test set are given in Table 1. The sentences of this test set contain context information for predicting the gender of the neutral word “friend” in English, either “amigo” or "amiga" in Spanish. The lower the bias in the system, the better the system will be able to translate the gender correctly. See Table 5 for the percentages of how "friend" is predicted for each model.

"Him" is predicted at almost 100% accuracy for all models. However not all occupations are well translated. On the other hand, the accuracy drops when predicting the word "her" on all models. When using names, the accuracy is even lower for "Mary" opposite to "John".

Note that gender debiasing is shown by augmenting the percentage of "amiga" in the translation in the presence of the female pronoun while keeping the quality of translation (coherently with generic results in Table 4). Based on accuracy values from Table 5, the most neutral system is achieved with GloVe and also with Hard-Debiased GloVe pre-trained embeddings. The accuracy improves by 30 percentage points compared to the baseline system and over 10 percentage points compared to the non-debiased pre-trained word embeddings.

The quality of the translation also depends on the professions from the *Occupations test* and its predicted gender. Again, the system has no problem predicting the gender of professions in the context of "him", so we focus the analysis on the context of "her". With GN-GloVe pre-trained embeddings both in the encoder and decoder sides, the model shows a higher accuracy when predicting the gender of a profession in Spanish. Specifically, for technical professions such as "criminal investigator", "heating mechanic", "refrigeration mechanic" and others such as "mine shuttle car operator". See Table 6 for the prediction on this last profession.

7 Conclusions and further work

Biases learned from human generated corpora is a topic that has gained relevance over the years. Specifically, for MT, studies quantifying gender bias present in news corpora and proposing debiasing approaches for word embedding models have shown improvements on this matter.

We studied the impact of gender debiasing on neural MT. We trained sets of word embeddings with the standard GloVe algorithm. Then, we debiased the embeddings using a post-process method (Bolukbasi et al., 2016) and also trained a gender neutral version (Zhao et al., 2018b). We used all these different models on the Transformer (Vaswani et al., 2017). Experiments were reported on using these embeddings on both the encoder and decoder sides, or only the encoder or the decoder sides.

The models were evaluated using the BLEU metric on the standard task of the WMT *newstest2013* test set. BLEU performance increase when using pre-trained word embeddings and it is slightly better for the debiased models.

In order to study the bias on the translations, we evaluate the systems on a custom test set composed of occupations. This set consists of sentences that include context of the gender of the ambiguous "friend" in the English-Spanish translation. This word can be translated to feminine or masculine and the proper translation has to be derived from context. We verified our hypothesis that consisted on the fact that if the translation system is gender biased, the context is disregarded, while if the system is neutral, the translation is correct (since it has the information of gender in the sentence). Results show that the male pronoun is always identified, despite not all occupations are well translated, while the female pronoun has different ratio of appearance for different models. In fact, the accuracy when predicting the gender for this test set is improved for some settings, when using the debiased and gender neutral word embeddings. Also, as mentioned, this system slightly improves the BLEU performance from the baseline translation system. Therefore, we are "equalizing" the translation, while keeping its quality. Experimental material from this paper is available online [7].

As far as we are concerned, this is one of the pioneer works on proposing gender debiased translation systems with word embedding techniques.

We did our study in the domain of news articles and professions. However, human corpora has a broad spectrum of categories, as an instance: industrial, medical, legal that may rise other biases particular to each area. Also, other language pairs with different degree in specifying gender information in their written or spoken communication could be studied for the evaluation of debiasing in MT. Furthermore, while we studied gender as a bias in MT, other social constructs and stereotypes may be present in corpora, whether individually or combined, such as race, religious beliefs

[7] https://github.com/joelescudefont/genbiasmt

Pre-trained embeddings	her amiga	him amigo	Mary amiga	John amigo
None	99.8	99.9	69.5	99.9
GloVe (Enc.)	2.6	100.0	0.0	100.0
GloVe (Dec.)	95.0	100.0	4.0	100.0
GloVe (Enc./Dec.)	**100.0**	100.0	90.0	100.0
GloVe Hard-Debiased (Enc.)	**100.0**	100.0	99.5	100.0
GloVe Hard-Debiased (Dec.)	12.0	100.0	0.0	100.0
GloVe Hard-Debiased (Enc./Dec.)	99.9	100.0	**100.0**	99.9
GN-GloVe (Enc.)	**100.0**	100.0	7.7	100.0
GN-GloVe (Dec.)	97.2	100.0	51.8	100.0
GN-GloVe (Enc./Dec.)	99.6	100.0	56.4	100.0

Table 5: Percentage of "friend" being translated as "amiga" or "amigo" in test sentences with female-male pronouns and proper names for the *Occupations test*. Best results in bold.

or age; this being just a small subset of possible biases which will present new challenges for fairness both in machine learning and MT.

Acknowledgments

This work is supported in part by the Spanish Ministerio de Economía y Competitividad, the European Regional Development Fund and the Agencia Estatal de Investigación, through the postdoctoral senior grant Ramón y Cajal, the contract TEC2015-69266-P (MINECO/FEDER,EU) and the contract PCIN-2017-079 (AEI/MINECO).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. *CoRR*, abs/1705.03122.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Nishtha Madaan, Sameep Mehta, Shravika Mittal, and Ashima Suvarna. 2018. Judging a book by its description : Analyzing gender stereotypes in the man bookers prize winning fiction. *CoRR*, abs/1807.10615.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. 2018. Assessing gender bias in machine translation - A case study with google translate. *CoRR*, abs/1809.02208.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? 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 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

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.

Pre-trained word embeddings	Prediction ***La*** *conozco desde hace mucho tiempo,*
None	*mi amigo trabaja como mecánico de refrigeración.*
GloVe (Enc)	*mi **amiga** trabaja como mecánico de refrigeración.*
GloVe (Dec)	*mi **amiga** trabaja como mecánico de refrigeración.*
GloVe (Enc+Dec)	*mi **amiga** trabaja como mecánico de refrigeración.*
GloVe Hard-Debiased (Enc)	*mi amigo trabaja como mecánico de refrigeración.*
GloVe Hard-Debiased (Dec)	*mi **amiga** trabaja como mecánico de refrigeración.*
GloVe Hard-Debiased (Enc+Dec)	*mi **amiga** trabaja como mecánico de refrigeración.*
GN-GloVe (Enc)	*mi **amiga** trabaja como mecánico de refrigeración.*
GN-GloVe (Dec)	*mi **amiga** trabaja como mecánico de refrigeración.*
GN-GloVe (Enc+Dec)	*mi **amiga** trabaja como **mecánica de refrigeración**.*
Reference	*mi amiga trabaja como mecánica de refrigeración.*

Pre-trained word embeddings	Prediction ***La*** *conozco desde hace mucho tiempo,*
None	*mi **amiga** trabaja como operador de un coche de enlace a las minas.*
GloVe (Enc)	*mi amigo trabaja como operador del transbordador espacial.*
GloVe (Dec)	*mi **amiga** trabaja como un operador de transporte de camiones.*
GloVe (Enc+Dec)	*mi **amiga** trabaja como un operator de coches.*
GloVe Hard-Debiased (Enc)	*mi **amiga** trabaja como mine de minas.*
GloVe Hard-Debiased (Dec)	*mi amigo trabaja como un operador de transporte de coches para las minas.*
GloVe Hard-Debiased (Enc+Dec)	*mi **amiga** trabaja como un operator de coches.*
GN-GloVe (Enc)	*mi **amiga** trabaja como operador de ómnibus de minas.*
GN-GloVe (Dec)	*mi **amiga** trabaja como un operador de transporte para las minas.*
GN-GloVe (Enc+Dec)	*mi **amiga** trabaja como **operadora de transporte de minas**.*
Reference	*mi amiga trabaja como operadora de vagones de minas.*

Table 6: Spanish predictions for the test sentences "*I've known her for a long time, my friend works as a refrigeration mechanic.*" "*I've known her for a long time, my friend works as a mine shuttle car operator.*". Best results in bold.

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. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, pages 2979–2989. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL-HLT (2)*, pages 15–20. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4847–4853.

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)*, Paris, France. European Language Resources Association (ELRA).