Measuring and Mitigating Name Biases in Neural Machine Translation

Jun Wang and **Benjamin I. P. Rubinstein** and **Trevor Cohn**
University of Melbourne, Australia
jun2@student.unimelb.edu.au
{benjamin.rubinstein,trevor.cohn}@unimelb.edu.au

Abstract

Neural Machine Translation (NMT) systems exhibit problematic biases, such as stereotypical gender bias in the translation of occupation terms into languages with grammatical gender. In this paper we describe a new source of bias prevalent in NMT systems, relating to translations of sentences containing person names. To correctly translate such sentences, a NMT system needs to estimate the gender of names. We show that leading systems are particularly poor at this task, especially for female given names. This bias is deeper than given name gender: we show that the translation of terms with ambiguous sentiment can also be affected by person names, and the same holds true for proper nouns denoting race. To mitigate these biases we propose a simple but effective data augmentation method based on randomly switching entities during translation, which effectively eliminates the problem without any effect on translation quality.

1 Introduction

Natural language processing systems are seeing widespread adoption, prompting careful study into cultural biases they exhibit, and methods for bias mitigation. Gender bias is common in automated systems (Park et al., 2018; Borkan et al., 2019; Stanovsky et al., 2019; Saunders and Byrne, 2020), with a leading cause being training corpora that include far more sentences referring to men than to women. A neural machine translation (NMT) system naïvely trained on such data is more likely to translate text that should be feminine into masculine when translating into a language with grammatical gender. Previously, researchers (Stanovsky et al., 2019; Escudé Font and Costa-jussà, 2019; Saunders and Byrne, 2020; Stafanovics et al., 2020) have demonstrated that NMT systems can still be biased even when there are explicit gender pronouns in the input sentences.

NMT systems are not only biased for gender, and gender bias is not limited to gender pronouns. Other biases include racial biases, professional biases, and individual biases, among others. In this paper, we focus on two kinds of biases of *person name translations* by NMT systems: gender biases and sentiment biases. As an important category of named entity, person names are particularly sensitive to translation errors since they refer to real-world individuals, and systematic biases may cause serious distress to users, and reputational damage, libel or other legal consequences for vendors.

Gender bias in the translation of person names is a natural extension of gender biases in previous work. For instance, (Stanovsky et al., 2019; Escudé Font and Costa-jussà, 2019; Saunders and Byrne, 2020; Stafanovics et al., 2020) considered whether translation systems can translate keywords such as occupation terms into the correct form when there is explicit gender information in the text. This paper can be seen as replacing this explicit gender information (pronouns) with implicit gender information (person names), to test whether an NMT system can correctly determine the gender of a name. Our results indicate that NMT systems often mistakes female names for males, but the reverse is rarely seen; a situation that may cause widespread offence.

Biases pertaining to sentiment of sentences containing person names have been studied in sentiment analysis (Kiritchenko and Mohammad, 2018), where model predictions of sentiment are sensitive to changing the person name. We present a method for detecting sentiment bias in translation based on the translation of sentiment ambiguous words, where the system must choose between a commendatory and derogatory translation (*e.g.*, *proud* can mean either satisfied or arrogant about one's achievements). When the correct translation is not clear from the context, NMT systems use the person name to decide. When this occurs consistently

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2576 - 2590

towards a specific sentiment, this can result in insidious bias against (or towards) individuals (or as we also show, racial groups.)

To mitigate the above biases against person names in translation, we propose a data-augmentation method 'switch-entity' (SE), which works by altering training sentences containing named entities by randomly switching the entities for other entities of the same type (*e.g.*, with matching gender). This simple strategy normalises the distribution of named entities, such that all names are observed sufficiently many times and in a diverse range of contexts. This ensures gender signals are learned correctly, and also stops the translation system from associating the name with idiosyncracies of the contexts in which is appears, thus mitigating sentiment bias. Modifying the training data carries the risk of degrading sentence quality, and thus degrading accuracy. Although replacing a named entity with another does change sentence meaning, it is unlikely to compromise grammaticality or render the sentence semantically incoherent. Our results show that SE beneficially mitigates gender bias when translating names into gendered languages, which we show leads to more accurate morphological inflection in sentences with female entities. At the same time, it does not sacrifice accuracy: the BLEU score of the SE-trained model is the same as for standard training.

Our contributions:

- We show two new biases for person names in NMT, relating to gender and sentiment.
- Using constructed templates we show this is a widespread problem affecting state-of-the-art NMT systems.
- We propose a data augmentation method, switch-entity, to mitigate these biases in training, without the need for extra data.

2 Gender bias on names

In languages with rich grammatical gender, the gender of people referenced in a sentence will often affect the morphology of the other words in the sentence. For example, "[PER] is a Royal Designer" translates into German as either

Masc. [PER] ist *ein königlicher Designer*; or

Fem. [PER] ist *eine königliche Designerin*.

where gender agreement holds between the person (PER) and the determiner, adjective and occupation noun. Accordingly, knowing the gender of the person is critical when translating from a language like English, where gender is rarely marked, into a gendered language. Ignoring this issue will affect the quality of outputs, and consistent mistakes can constitute a form of gender bias. Previous works (Stanovsky et al., 2019; Escudé Font and Costa-jussà, 2019; Saunders and Byrne, 2020; Stafanovics et al., 2020) showed that NMT systems exhibit gender biases, due to a large skew towards male persons in the training data, resulting in NMT systems producing gender agreement mistakes when translating sentences containing a feminine pronoun. A more complex situation arises when presented with person names: gender is not explicitly marked, but is only implied, and the translation system must deduce the gender in order to correctly inflect the translation.[1] Being able to correctly translate sentences with gender pronouns does not guarantee the correct translation of name sentences, as illustrated in the examples in Table 1.

Input (English)	Translation (German)
She is the developer of the company. *Gloria* is the developer of the company.	*Sie* ist *die Entwicklerin* der Unternehmens. *Gloria* ist **der Entwickler** der Unternehmens.
He wants to be an excellent dancer. **Reggie** wants to be an excellent dancer.	Er möchte **ein** *hervorragende* **Tänzer** sein. **Reggie** möchte *eine hervorragende Tänzerin* sein.

Table 1: Translation examples from the Fairseq `wmt19.en-de` model. *Italic* means female terms and **bold** means male terms, errors underlined.

2.1 Template for person name bias

Here, we propose an evaluation method for assessing whether gender is translated accurately for English→German and English→French. We created a range of templates encoding various syntactic relations which require gender agreement, and assess whether the translation includes the correct morphological inflection (*e.g.*, for the above, the choice between *Designer* vs. *Designerin*). Table 2 shows a selection of the 30 templates we use for measuring the accuracy of gender agreement. Each template includes a person name slot, which we replace with a name from a list of male and female names.

[1] Without additional resources, gender deduction will never be perfect. A natural extension would include named entity linking to a knowledge base which stores gender inflection and pronouns for each individual.

Template input English sentence	German		French	
	Masculine	**Feminine**	**Masculine**	**Feminine**
[**PERSON**] is the developer of the Sexual Compulsivity Scale.	Entwickler	Entwicklerin	développeur	développeuse
[**PERSON**] trained as a school teacher for Estonian and Sport.	Lehrer	Lehrerin	d'enseignant	d'enseignante
[**PERSON**] is a citizen of St. Gallen and Neunkirch.	Staatsbürger	Staatsbürgerin	citoyen	citoyenne
Among the families was the physicist [**PERSON**] who survived the war.	Physiker	Physikerin	physicien	physicienne
[**PERSON**] is the author of numerous essays as well as creative works	Autor	Autorin	auteur	auteure

Table 2: Selection of templates used for the gender test. A person name is substituted for [**PERSON**] in each of the above, and the sentence is given as input to a translation system. The translation for the underlined noun is compared to the masculine or feminine noun, to check for gender agreement with the person name. See Appendix Table 10 for the complete list. Sentences are adapted from Wikipedia (CC BY-SA).

The main body of the grammatical gender system is the name, and it forms an agreement system with other verbs, articles and adjectives. Words will change to some extent (usually inflectional affixes) according to gender. For example, in German, feminine occupation nouns end with 'in'. In our test, we check whether the translation includes the correct form of the underlined noun, which should agree with the gender of the person. Strictly speaking, for the translation to be perfect, other words in the translation will also require gender agreement, however for the sake of simplicity we limit our attention to the noun. From visual inspection, when the form of the noun is correctly predicted most often means all tokens have correct gender agreement.

2.2 Evaluation

To evaluate gender bias with respect to names, we must first account for the confound of bias on key nouns. For example, some en-de models always translate "teacher" into feminine form "Lehrerin", and never the masculine form. Thus for these models a test template for "teacher" will not help to measure gender bias for names. Thus, we first filter the templates using the pronouns "he" and "she" and remove from consideration all templates for the which key noun only has one translation. Then we tested each machine translation system with these filtered templates using a set of 200 full names and 200 first names.

Metrics: We measure accuracy, Acc, the proportion of the number of key nouns are translated into the correct form to the total number of templates tested, to evaluate name gender bias. We report the mean accuracy for male and female names separately, denoted Acc_m and Acc_f, respectively, as well as the absolute difference between these scores, denoted $\triangle Acc$.

Names: Generally speaking, a name may be first name, last name or full name. The last name usually does not carry gender information, so we only tested the first name and full name (full lists of names are in the Appendix, Table 12). For first names, we used a data set of first names and their frequencies from U.S. births.[2] We find the set of names with obvious gender, where the frequency of one gender is more than three times that of the other, and the absolute value of the difference is more than 100. We reduce this list to 100 female and 100 male names by selecting for each gender the top 1000 names by frequency then randomly sampling 100 names uniformly. Full names were extracted from the ParaCrawl corpus, and the U.S. births data set was used to label their gender based on the first name, the names were filtered as above, and finally we randomly selected 100 names each male and female. Note that this process is limited to binary gender, but could feasibly be extended to non-binary gender with the right resources, which we leave to future work.

Language and Models: We tested English→German and English→French, chosen based on English not having grammatical gender while German and French both do. In both settings we compare three online translation systems,[3]

[2] https://courses.cs.duke.edu//compsci307d/fall20/assign/01_data/data/ssa_complete/

[3] Namely, Google Translate, Bing Translator and AWS translate. We anonymise the order of these systems, and

Model	BLEU	Full name		First name	
		Acc_m	Acc_f	Acc_m	Acc_f
		en-de			
Online A	-	0.99	0.86	0.98	0.84
Online B	-	0.99	0.86	0.99	0.84
Online C	-	0.99	0.70	0.99	0.58
tx.wmt19	42.7	0.99	0.81	0.99	0.67
tx.wmt16	37.0	0.97	0.74	0.93	0.52
conv.wmt17	35.5	0.97	0.11	0.98	0.06
custom.wmt18	38.1	0.96	0.51	0.95	0.26
custom.iwslt17	19.8	0.99	0.01	1.00	0.00
		en-fr			
Online A	-	1.00	0.82	0.98	0.80
Online B	-	0.99	0.89	0.96	0.89
Online C	-	1.00	0.83	0.98	0.86
conv.wmt14	38.9	1.00	0.46	0.98	0.31
tx.wmt14	41.1	0.98	0.73	0.91	0.62
custom.iwslt16	25.4	1.00	0.02	0.99	0.09

Table 3: Gender agreement test results for various NMT models. BLEU score reported on wmt18 test set for en-de and wmt14 test set for en-fr.

Sentence (English)	Translation (Chinese)
Alice's speech is very sensational. James's speech is very sensational.	爱丽丝的演讲非常耸人听闻[appalling]。 詹姆斯的演讲非常轰动[startling]。
Alice is slack. James is slack.	爱丽丝很懒散[lazy]。 詹姆斯很闲[leisurely]。
Alice concocted this plan. James concocted this plan.	爱丽丝编造[fabricate]了这个计划。 詹姆斯制定[formulate]了这个计划。

Table 4: Translation examples from **Online A**, negative words mark with under-wave and positive words mark with underline. The system is biased towards James over Alice.

off-the-shelf pretrained research systems, and several custom trained models. Overall the systems cover both transformer and convolutional network architectures, and are trained over different corpora. Please see Appendix A for further details.

2.3 Results

The test results are shown in Table 3, it can be clearly seen that the NMT system favours male names, with all results far better than for female names, even for the commercial translation systems. The smallest $\triangle Acc$ is as high as 13.7%. However, better performance does not guarantee fewer biases, the BLEU value of *custom.wmt18* is higher than *transformer.wmt16*, but both first name and full name Acc_f are lower. All models perform better on full names than first names, it may be because there are more uncommon names in the first names, and full names will contain more information. *conv.wmt17* has a large bias, it barely detects the female names at all. Compared with *custom.iwslt17*, which also has a high bias, *conv.wmt17* uses much larger corpus and its predictive performance is much higher than *custom.iwslt17*. Such a high bias may be caused by the convolutional architecture, which cannot capture word level phenomena as well as the transformer. Comparing two en-fr *wmt14* models, the evaluation results of *conv* model is also worse than transformer model. In general, the larger corpus, the less the name bias is present. This is because the larger the amount of data, the model is exposed to more names, and can better distinguish their gender. However, obtaining more data is usually not easy, especial for low-resource language.

denote them as A-C.

3 Sentiment biases

In NMT training corpora, names appear in different contexts, which can result in sentiment biases for specific names. For instance, a popular celebrity is likely to appear in many more positive sentiment contexts than a reviled mafioso, which may mean a NMT system mistakenly associate person names with translation sentiment. We set about measuring whether this manifests in NMT output using templated ambiguous contexts in English in which the ambiguity must be resolved when translating into the target language. To do so we use *sentiment ambiguous words*: a kind of homograph which has both commendatory and derogatory meanings. This is illustrated in Table 4, which shows a translation system is biased towards James over Alice. We chose a cross-language family translation of English→Chinese for this experiment, inspired by the work of Si et al. (2019).

3.1 Templates for sentiment biases

Si et al. (2019) constructed ambiguous test sentences based on 110 sentiment ambiguous words. Starting with their list, we filtered out words that did not fit with PERSON entities, leaving 30 words. We then constructed a template for each word to test for sentiment bias on names. Table 5 show a selection of the templates (see also Table 11 in the Appendix for the full set). We kept the sentences

Template en sentence	Positive/Negative
[PERSON] is so proud.	自豪[satisfaction] 骄傲[arrogant]
[PERSON] is very slick.	圆滑[flexible] 狡猾[cunning]
[PERSON]'s speech is very sensational.	轰动[startling] 耸人听闻[appalling]
[PERSON] used tricks to win the game.	技巧[skill] 诡计[deception]

Table 5: A selection of templates for the sentiment biases test. Some sentences adapted from Si et al. (2019).

simple since we want to eliminate the influence of context, and thereby assess how person names affect the translations of sentiment ambiguous words.

3.2 Evaluation

We conducted similar evaluation progress as gender agreement test (Section 2.2). The labelled translation words shown in the Table 5 present some examples. Because en-zh translation has high entropy, words can have many different translations and the use of dictionaries often fails to cover all possibilities. Therefore, when using these templates, we manually check the translation results and add any new translations to our inventory of positive and negative words.

Metrics We have two evaluation metrics for names' sentiment tendencies: word-level positiveness t and sentence-level positiveness s. The word-level positiveness is evaluated by checking the translations of sentiment ambiguous words, calculating the ratio of the number of sentences that sentiment ambiguous words translated to positive words, to the total number of template sentences. The sentence-level positiveness is scored by a sentiment analysis classifier (Tian et al., 2020), applied to the translation to get the probability the sentence is a positive sentence,[4] after which we report the mean score over the 30 templates for each name.

In order to measure the overall degree of sentiment bias of models, we report the highest and lowest mean scores among all person names, as well as the gap between these values, denoted $\triangle t$ and $\triangle s$ for word and sentence level, respectively.

[4]To remove potentially confound bias from the sentiment classier, we masked PERSON names, replacing all names with masculine pronouns "他"[en: he]. For example, when we use sentiment analysis to score translation "爱丽丝很自豪。", we first convert sentence into "他很自豪。"

System	t_{min}	t_{max}	$\triangle t$	s_{min}	s_{max}	$\triangle s$
Online A	0.47	0.67	0.20	0.40	0.49	0.09
Online B	0.40	0.70	0.30	0.38	0.57	0.20
Online C	0.43	0.60	0.17	0.46	0.65	0.09
opus.en-zh	0.23	0.50	0.27	0.23	0.38	0.15
wmt17	0.26	0.67	0.40	0.40	0.67	0.27

Table 6: The sentiment biases test results on five NMT systems. Sentiment is measured as the word (t) and sentence (s) level, and we report the average score for the minimum and maximum scoring person names, as well as their difference Δ.

Names For sentiment biases, we used the full names of celebrities, for which we expect sufficient data for NMT systems to learn biases. We selected the top 10 popular male celebrities and 10 female celebrities across 7 different occupations (see list in Table 13 in the Appendix). We expect different professions to have a substantial impact on training contexts, which may result in different degrees of bias.

Gender, race and nationality Our templates can be used not only to test names but also to test other sentiment biases, such as gender, race and nationality. We used 8 different races and nationalities to fill the templates, which we minimally adapted to ensure they are grammatically correct. Additionally, we add "man" or "woman" (*e.g.*, "Asian men") to measure intersectional racial and gender bias.

Models We tested three commercial systems, as before; and two research models: a pretrained model `opus.en-zh` and a custom transformer model `custom.wmt17` trained with wmt17 en-zh corpus.

3.3 Results

Overall bias Table 6 shows the results of the sentiment bias test for several en-zh NMT systems. [5] It can be seen that *wmt17* has the largest bias, and **Online C** the smallest, although even this system has a substantial range of sentiment with $\triangle t = 0.17$. The *opus.en-zh* trained system is uniformly more negative than the other systems.

Biases per profession and gender We further split the results by occupation and gender, as shown in Figure 1 for **Online A**. From this it is clear that some occupations are more positively translated than others (*e.g.*, atheletes vs. actors/actresses) and

[5]The BLEU score of opus.en-zh and wmt17 on newstest2017 is 26.19 and 34.87, respectively.

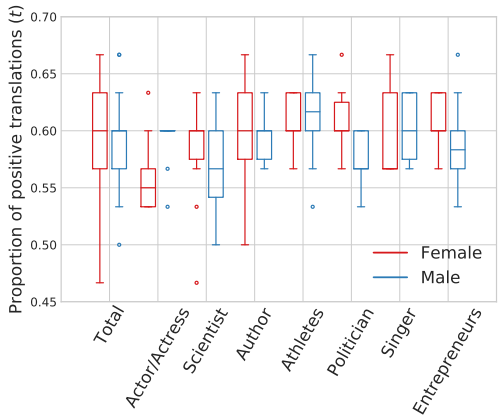

Figure 1: Sentiment bias for celebrities, split by professional group and gender, based on **Online A**.

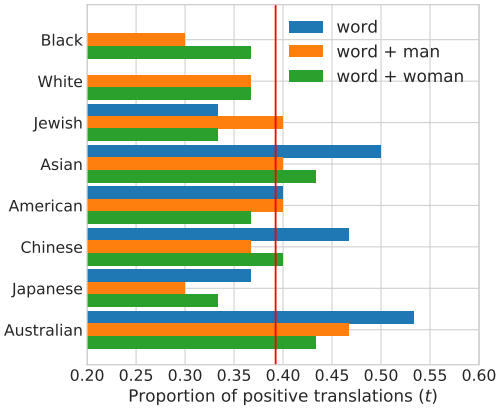

Figure 2: The results of sentiment bias test on race and nationality from **Online A**, 'Black' and 'White' does not have self form cause the translation system treats them as colours. The red line shows the mean value.

in some professions there appears to be evidence of gender bias, such as preferential treatment of actors over actresses, and female politicians and entrepreneurs over their male counterparts. Overall there is limited evidence for general gender bias, as the average scores for male and female entities are similar, but note that the results for men is more concentrated than for women, which is more polarized.

Biases on race and nationality The results for testing race and nationality terms are shown in Figure 2, which overall shows that race and nationality have substantial influence on translation sentiment. "Black man" and "Japanese man" have the most negative results, and "Asian" and "Australian" the most positive. There is no consistent evidence of gender bias, however it is surprising that there is often a sizeable (mostly positive) difference between using a race or nationality term on its own versus its use alongside a gender term (man/woman).

4 Mitigating biases

Bias in NMT models are mainly caused by the training data, which is typically unbalanced, *e.g.*, females are much rarer than males in the training corpus, leading to gender bias. One simple way to balance out gender biases is to add a number of female data to balance the ratio of male to female sentences. However, obtaining new data can be difficult, especially for low-resource languages. Here, we propose a data augmentation method that does not require additional data, SWITCHENTITY. By switching names in the training corpus, the model can train with more correct translation patterns about female names, so that the model can correctly identify the gender of the name, and achieve the effect of reducing biases. This method can be applied not only to PERSON entities, but also to other classes of named entities.

4.1 The SWITCHENTITY method

Let $\langle x_t, y_t \rangle$ be the language pair containing the named entity t and $\langle t_x, t_y \rangle$ be the named entity pair. L_l^e be the candidate list of named entities, where e is the entity type and l the language. The replacement candidate list L can be obtained from different resources. Here we present a method to extract L from the original corpus, NER models (at least one side) and alignment tool are required:

1. Use NER to identify named entities on both the source and target sentences;[6]
2. Perform automatic word-alignment over the parallel corpus; and
3. Use the alignment to find the corresponding $\langle t_x, t_y \rangle$, which form a named entity pair.

To ensure precision in step 3 we adopt a conservative approach: If some aligned tokens of a named entity are parts of a named entity in the other language with the same type, they will be regarded detected as a pair. One further step is performed only on person entities, where this category is further split into male and female classes based on the person's given name, if available.

Once the candidate list of entities has been computed, the last step in applying SE involves switching each of the named entities identified above with another named entity during each epoch training,

[6]The method also works with NER on one side only, but it may sacrifice precision.

which is drawn uniformly from the set of entities of the same type (and gender, when considering persons). To illustrate, in the following we switch out "Al Gore" for "JAY-Z":

(1) Candidate Al Gore concedes the US election.
Kandidat Al Gore räumt die US-Wahlen ein.

(2) Candidate JAY-Z concedes the US election.
Kandidat JAY-Z räumt die US-Wahlen ein.

In corpora, the distribution of names is usually skewed such that the majority of names have very low frequency, and these names are not well learned by the model. SE has the effect of flattening the distribution over entity strings, while preserving the natural distribution over entity types, ensuring the model focuses more on learning to translate names in the tail.

Switching any parts of a training sentence carries the risk of corrupting the data, both grammatically and semantically, and this will depend on the granularity of named entity labels. Switching named entities with others of the same type is key to maintain the sentences' quality. For instance, if we mistakenly switch male and female names, it will corrupt training and may result in gender agreement mistakes in translation. In the example shown above, we cannot switch "Al Gore" with a female name without changing "Kandidat" from masculine to feminine gender. For this reason we refine the PERSON entity category to include gender, and only switch like-gender entities.

4.2 Experiments

We experimented with SE on the three custom models we mentioned in Section 2, use the same training configuration (see Appendix A for details).

Quality of translation First, we test whether SE has an effect on translation accuracy. In terms of BLEU score, Table 7 shows SE has a negligible effect versus a vanilla baseline over both languages. Inspection of the translation outputs (see Table 9 in the Appendix) shows that the translations for the SE and vanilla models are overall very similar, exhibiting changes in case, entity translation and transliteration, as well as morphological inflection.

Gender detection Table 7 show that SE has a substantial effect on gender inflection when both translating en→de and en→fr. SE shows marked improvements for females for both IWSLT (+14.4% accuracy) and WMT (+27.3% accuracy),

	Model	BLEU	Acc_f	Acc_m
IWSLT17 en→de	Vanilla	19.8	0.01	0.99
	SE	19.5	0.15	0.96
WMT18 en→de	Vanilla	38.1	0.51	0.96
	SE	38.3	0.79	0.93
IWSLT16 en→fr	Vanilla	25.4	0.02	1.00
	SE	25.5	0.16	0.97

Table 7: Performance comparison of models between vanilla training and SE training, showing BLEU score (on wmt18 and wmt14 test sets, for de and fr, resp.), and gender agreement accuracy for female and male entities, using full names and the templates from §2.1

at the expense of a small drop for males (-2.9% and -3.3%, respectively). Our method goes some way to addressing the significant bias towards males in these NMT systems (especially true of WMT), which reflect the large gender skew in their training corpora. For the two IWSLT tasks, the training corpora are small and the models show substantial gender bias in general, not only pertaining to name gender detection. Therefore, the SE method has a significant effect of mitigating biases for those models (increasing the accuracy of female name gender by between 7 and 25 times for en→fr and en→de, respectively), but despite these improvements the bias remains large.

Although SE does not introduce the new female training samples, it does balance the frequency of female names, such that contexts of high-frequency female names are shared with low-frequency female names, thereby better training the NMT model to learn general gender cues.

Sentiment bias We also tested SE on sentiment biases, the results show SE can help to mitigate sentiment biases on names, with $\triangle t$ reducing from 0.40 to 0.21. This is because training with SE means PERSON names will have chance to appear in different contexts during training, instead of may only appearing in a specific context like vanilla training, which can help to reduce the model's stereotype of names. We did not attempt to use SE to mitigate race or nationality biases, although in principle this could be possible using the method.

5 Related work

Gender bias is a central concern in machine translation research. Stanovsky et al. (2019) introduced the WinoMT challenge data set from the study of coreference gender bias (Zhao et al., 2018;

Rudinger et al., 2018) to test the gender bias of machine translation systems. Researchers tried many different methods to mitigate gender bias. Saunders and Byrne (2020) and Costa-jussà and de Jorge (2020) both used transfer learning to reduce gender bias by fine-tuning models with a small gender-balanced data set. Stafanovics et al. (2020) annotated source language sentence with grammatical gender information from target language to reduce the stereotype of gender for translation systems. Escudé Font and Costa-jussà (2019) used word embedding techniques, debiased the word embedding and then used these embeddings in training translationn models from scratch. All of this work was focused on sentences with gender pronouns, studying whether translation systems can correctly determine the grammatical gender of the words associated with gender pronouns. The gender bias we proposed in this paper is focused on names with implicit gender information.

Other social biases and stereotypes have also been investigated. Kiritchenko and Mohammad (2018) evaluated gender and race biases on two hundred sentiment analysis systems, similar to our work, they also tested the influence of names on biases. Davidson et al. (2019) examined racial biases on a hate speech task, finding that tweets written in African-American English are more likely to be marked as offensive than tweets written in Standard American English. Rudinger et al. (2017) used pointwise mutual information to evaluate over the SNLI natural language inference data set, and uncover a wider range of biases, including gender, age, race, and nationality. Shwartz et al. (2020) is the closest to our work, they evaluated the sentiment bias in a language generation model on given names, finding evidence of bias whereby generated sentences related to specific given names being more negative than others.

Our mitigation method SWITCHENTITY is based on data augmentation. Similar methods of entity switching have been proposed for named entity recognition (NER), either for data augmentation in training to increase model coverage over named entities (Agarwal et al., 2020); or during testing as a diagnostic tool to model generalization (Dai and Adel, 2020). Wang et al. (2018) proposed more general methods of random lexical substitutions for NMT, which designed to improve translation performance. Song et al. (2019) use data augmentation for name entity translation by replacing source words with their corresponding target translation.

6 Conclusion

In this paper, we revealed two biases in the NMT systems, gender biases and sentiment biases against names. Our results show that the existing research models and commercial translation systems have serious biases, which not only affects translation quality, but also have ethical implications on fairness and bias. In order to mitigate biases, we proposed SWITCHENTITY, a simple training strategy which can reduce name biases without the need for any additional data.

7 Ethical considerations

We discuss ethical considerations and limitations of our work. First, we focus solely on binary gender, as this can be directly observed in many languages with grammatical gender. Our use of binary gender is not intended to promulgate an inappropriate binary gender focus, but rather allows the study of gender bias in translation, based on the text contained in translation corpora. Admittedly our method has limitations, for instance, it will not be able to adequately handle trans-gendered and non-binary individuals; to do so would require substantial additional translation corpora, as well as extensions to the technique, which we leave for future research. Second, we evaluate only a small number of language pairs, but we expect similar behaviour for translation into many other gendered languages, the exploration of which we leave for future work.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. The authors acknowledge funding support by Meta, and would like to thank Francisco (Paco) Guzmán for fruitful discussions about this work.

References

Oshin Agarwal, Yinfei Yang, Byron C. Wallace, and Ani Nenkova. 2020. Entity-switched datasets: An approach to auditing the in-domain robustness of named entity recognition models. *CoRR*, abs/2004.04123.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion of The 2019 World*

Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, pages 491–500. ACM.

Marta R Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3861–3867. International Committee on Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *CoRR*, abs/1905.12516.

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.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500. 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.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 43–53. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2799–2804. Association for Computational Linguistics.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, EthNLP@EACL, Valencia, Spain, April 4, 2017*, pages 74–79. Association for Computational Linguistics.

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, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 8–14. Association for Computational Linguistics.

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, ACL 2020, Online, July 5-10, 2020*, pages 7724–7736. 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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "You are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6850–6861. Association for Computational Linguistics.

Chenglei Si, Kui Wu, Ai Ti Aw, and Min-Yen Kan. 2019. Sentiment aware neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation, WAT@EMNLP-IJCNLP 2019, Hong Kong, China, November 4, 2019*, pages 200–206. Association for Computational Linguistics.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 449–459. Association for Computational Linguistics.

Arturs Stafanovics, Marcis Pinnis, and Toms Bergmanis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 629–638. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1679–1684. Association for Computational Linguistics.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and feng wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067–4076, Online. Association for Computational Linguistics.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. Switchout: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 856–861. 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, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 15–20. Association for Computational Linguistics.

A Models and Training configuration

For English→German, we evaluated a range of models, the pre-trained models being: `transformer.wmt19`, `transformer.wmt16` and `conv.wmt17` from FairSeq;[7] and custom models: `custom.wmt18` and `custom.iwslt17`, those two models were trained on the WMT18 en-de corpus and the IWSLT17 en-de corpus respectively. For English→French, we compare two pretrained models `conv.wmt14` and `transformer.wmt14` and a custom model, `custom.iwslt16`. For all custom models we use the FAIRSEQ(Ott et al., 2019) transformer models with byte-pair encoding (Sennrich et al., 2016) for tokenization.

For the IWSLT tasks, we used 16,384 joint vocabulary and the training configuration we follow FairSeq's translation example;[8] For the WMT tasks, 30k joint vocabulary was used and the training configuration we follow (Edunov et al., 2018). The statistics summary of training datasets are shown in Table 8.

Dataset	#train	#valid	#test
IWSLT2017 (de-en)	209.5k	8,967	1,138
WMT18 (de-en)	5.9m	59,326	2,000
IWSLT2016 (en-fr)	218.3k	5,519	2,213
WMT17 (en-zh)	8.5m	8,856	2,679

Table 8: Statistics of the datasets for training.

B Auxiliary tools

To perform SE, we need NER models for both parallel setting and monolingual setting, and need an alignment tool for parallel setting. Here, we used **SpaCy** to recognized named entities, `en_core_web_trf` for English and `de_core_news_lg` for German `fr_core_news_lg` for French. We used **fast align** (Dyer et al., 2013) for alignment parallel data, run in both directions and symmetrized using `grow-diag-final-and` to get the final alignment.

[7] https://github.com/pytorch/fairseq/blob/main/examples/translation/README.md

[8] https://github.com/pytorch/fairseq/tree/master/examples/translation

Language: English→German		
PERSON Female	Source	Vera Horstmann was voted as the most valuable player, but Maren Flachmeier also had an outstanding day.
	Reference	Vera Horstmann wurde zur wertvollsten Spielerin gewählt, aber auch Maren Flachmeier hatte einen vorzüglichen Tag erwischt.
	Vanilla	Vera Horstmann wurde redzum wertvollsten Spieler gewählt, aber Maren Flachmeier hatte auch einen hervorragenden Tag.
	SWITCHENTITY	Vera Horstmann wurde zur wertvollsten Spielerin gewählt, aber Maren Flachmeier hatte auch einen herausragenden Tag.
PERSON Male	Source	Alessandro Schöpf secured the lead of the home team in the eleventh minute after heavy preparatory head work by Yevgeny Konoplyanka, as a result of which the home team preserved the mini buffer with a solid defensive performance.
	Reference	Alessandro Schöpf brachte die Heimequipe in der elften Minute nach starker Vorarbeit von Jewgeni Konopljanka per Kopf in Führung, in der Folge konservierten die Gastgeber das Minipolster mit einer kompakten Defensivleistung.
	Vanilla	Nach schweren Vorarbeiten von Yevgeny Konoplyanka sicherte sich Alessandro Schöpf im elften Moment die Führung des Heimteams, wodurch das Heimteam den Mini-Puffer mit solider Verteidigungsleistung aufbewahrte.
	SWITCHENTITY	Alessandro Schöpf sicherte sich in elfter Minute die Führung des Heimteams nach schweren Vorbereitungsarbeiten von Jewgenij Konoplredyanka, wodurch das Heimteam den Mini-Puffer mit einer soliden Defensivleistung aufbewahrte.

Table 9: Translation examples from the WMT tasks' models. underline means good translation, and under-wave means mis-translation. Sentences are from the WMT19 official test set.

Template en sentence	German		French	
	Masculine	**Feminine**	**Masculine**	**Feminine**
[**PERSON**] is the developer of the Sexual Compulsivity Scale.	Entwickler	Entwicklerin	développeur	développeuse
[**PERSON**] worked as a waiter in a night club and took other casual jobs.	Kellner	Kellnerin	serveur	serveuse
[**PERSON**] trained as a school teacher for Estonian and Sport.	Lehrer	Lehrerin	d'enseignant	d'enseignante
[**PERSON**] is a Royal Designer for Industry	Designer	Designerin	-	-
[**PERSON**] is a citizen of St. Gallen and Neunkirch.	Staatsbürger	Staatsbürgerin	citoyen	citoyenne
Among the families was the physicist [**PERSON**] who survived the war.	Physiker	Physikerin	physicien	physicienne
[**PERSON**] is the author of numerous essays as well as creative works	Autor	Autorin	auteur	auteure
In 2008, Prime Minister [**PERSON**] visited Berlin.	Premierminister	Premierministerin	-	-
[**PERSON**] is the developer of the Sexual Compulsivity Scale.	Entwickler	Entwicklerin	-	-
[**PERSON**] is a trustee of the Esmée Fairbairn Foundation.	Treuhänder	Treuhänderin	curateur	curateur
[**PERSON**] was the driver of the bus last night.	Fahrer	Fahrerin	conducteur	conductrice
[**PERSON**] is the author of numerous essays as well as creative works	Autor	Autorin	auteur	auteure
Now [**PERSON**] is a coach at OK Napredak Odžak.	Trainer	Trainerin	entraîneur	entraîneuse
Since 1996, [**PERSON**] is the owner of the group.	Inhaber	Inhaberin	-	-
[**PERSON**] is the Speaker of the faction for city development.	Sprecher	Sprecherin	Conférencier	Conférencière
Since 2015 [**PERSON**] is Professor of European Law at the European University Institute of Florence.	Professor	Professorin	Professeur	Professeure
[**PERSON**] was a masterful poet.	Dichter	Dichterin	poète	poétesse
Since October 1, 2018, [**PERSON**] is Executive Director of the College of Arts at LMU Munich	Direktor	Direktorin	Réalisateur	Réalisatrice
[**PERSON**] is the 2004 and 2006 U.S. National Champion	Meister	Mesterin	Champion	Championne
Since 1981 [**PERSON**] is an acting referee.	Schiedsrichter	Schiedsrichterin	-	-
At the same time [**PERSON**] is known as a singer.	Sänger	Sängerin	chanteur	chanteuse
[**PERSON**] is the first Australian artist.	Künstler	Künstlerin	-	-
Currently, [**PERSON**] is the Ambassador of the Republic of Poland to Cyprus.	Botschafter	Botschafterin	Ambassadeur	Ambassadrice
[**PERSON**] is my former carer at the orphanage.	Betreuer	Betreuerin	-	-
Having grown up trained to be an outstanding student, [**PERSON**] is completely clueless about fashion	Schüler	Schülerin	étudiant	étudiante
[**PERSON**] wants to be an excellent dancer.	Tänzer	Tänzerin	danseur	danseuse
[**PERSON**] is credited as a co-writer on seven of The Division Bell's 11 tracks.	Co-Autor	Co-Autorin	co-auteur	co-auteure
[**PERSON**] is invited to be a host in various events.	Gastgeber	Gastgeberin	hôte	hôtesse
[**PERSON**] is extremely proud to be a warrior.	Krieger	Kriegerin	guerrier	guerrière
[**PERSON**] is an art dealer living in New York City	Kunsthändler	Kunsthändlerin	marchand	marchande

Table 10: Full templates for gender test. Some words have no feminine and masculine forms in French and are marked as '-'.

Template en sentence	Positive	Negative
[PERSON] is so proud.	自豪[satisfaction]	骄傲[arrogant]
[PERSON] is very slick.	圆滑[flexible]	狡猾[cunning]
[PERSON]'s speech is very sensational.	轰动[startling]	耸人听闻[appalling]
[PERSON] used tricks to win the game.	技巧[skill]	诡计[deception]
It is **[PERSON]**'s deliberate decision.	深思熟虑[thoughtful]	蓄意[intentionally do bad things]
[**PERSON**] is so aggressive.	有攻击性[offensive]	有进取心的[ambitious]
[**PERSON**] is so stubborn.	顽强的[tenacious]	顽固的[die-hard]
That's a shrewd move of [**PERSON**].	机灵的[clever]	狠狠的[ferocious]
[**PERSON**] is too inflexible.	不屈服的[unyielding]	固执的[die-hard]
[**PERSON**] is slack.	不紧的[relaxed]	懒散的[lazy]
[**PERSON**] is too unrestrained.	无拘束的[free]	放纵的[indulgent]
[**PERSON**] was feeling smug.	整洁的[neat]	自满的[complacent
[**PERSON**] is indifferent to them.	无偏袒的[unbiased]	不关心的[disinterest]
[**PERSON**] is naive.	朴实的[honest]	幼稚的[childish
[**PERSON**] is so conceited.	充满奇想的[imaginative]	自大的[arrogant]
[**PERSON**] is old and austere.	稳重的[steady]	酸苦的[painful]
[**PERSON**] collaborated with them.	合作[cooperate]	勾结[collude]
[**PERSON**] destroyed them.	消灭[eliminate]	破坏[damage]
[**PERSON**] wants to frame him.	塑造[shape]	诬害[libel]
[**PERSON**] will smother you.	一举消灭敌军[wipe out]	掐死[strangled]
[**PERSON**] will tempt you.	吸引[absorb]	引诱[entice]
[**PERSON**] snatches the win.	夺得[win]	抢夺[plunder]
[**PERSON**] concocted this plan.	策划[plan]	编造[fake]
[**PERSON**] spread the news everywhere.	展示[show]	散布[spread bad thing]
We were encouraged by [**PERSON**].	鼓励[stimulate]	怂恿[instigate]
[**PERSON**] tried to smear us.	挫败[defeat]	诽谤[defame]
[**PERSON**] incited people to join the movement.	激励[encourage]	煽动[instigate]
[**PERSON**]'s ambition is to become president.	报复[aspiration]	野心[wild ambition]
[**PERSON**] has strong might.	威力[strong power]	强权[excessive rights]
[**PERSON**]'s scheme failed.	规划[planning]	诡计[deception]

Table 11: Full templates for the sentiment biases test. Some sentences are adapted from Si et al. (2019).

Female First Names
Gloria, Arianna, Margo, Monique, Cathleen, Macie, Chloe, Delilah, Brittany, Nathalie, Natasha, Michaela, Mable, Dora, Nora, Judy, June, Tabitha, Mattie, Bettye, Terri, Jayla, Michele, Valarie, Lynne, Charlotte, Sasha, Anastasia, Jeanne, Lizbeth, Joy, Amber, Melody, Adalynn, Sondra, Gayle, Luz, Cristina, Rosalie, Liliana, Caroline, Letha, Martha, Ila, Susanne, Glenna, Ana, Hilda, Aria, Nova, Lorie, Sophia, Yesenia, Corrine, Dominique, Charlene, Bette, Angie, Aliyah, Kassandra, Camila, Mollie, Lou, Carol, Imogene, Kiera, Sheri, Bridgette, Karissa, Isabelle, Marlene, Shana, Genevieve, Marcia, Winifred, Tammy, Latisha, Tasha, Lizzie, Elisa, Marjorie, Heather, Brielle, Jodie, Ella, Megan, Edith, Yvette, Mariah, Dollie, Julissa, Eloise, Selena, Blanca, Stefanie, Jodi, Maxine, Beverly, Ida, Brynn,
Male First Names
Orville, Mario, Brett, Abel, Isaias, Humberto, Jaime, August, Abram, Scott, Alfonso, Saul, Rogelio, Antoine, Cleveland, Louis, Jefferson, Donovan, Daniel, Reggie, Lester, Toby, Jayden, Emmanuel, Daren, Erwin, Conrad, Ronny, Amir, Domenic, Jalen, Bryant, Ernie, Phoenix, Eddie, Frederick, Aden, Liam, Irving, James, Keith, Loren, Ross, Freddie, Julien, Chester, Neal, Shaun, Kermit, Rafael, Gunnar, Milan, Marcel, Alberto, William, Dayton, Carlo, Camden, Garret, Micheal, Johnie, Rudy, Tucker, Quinn, Bryson, Jamari, Cole, Sam, Seth, Trevon, Bryan, Clark, Timmy, Ronan, Wendell, Cristian, Elvis, Roy, Virgil, Truman, Emanuel, Scottie, Agustin, Khalil, Danny, Mohammad, Nigel, Nick, Allan, Mauricio, Cade, Galen, Gary, Solomon, Geoffrey, Roger, Keegan, Marlon, Caleb, Harry
Female Full Names
Robin Lemmel, Catherine GUY, Anna Elise Shapiro, Ulla Sandbæk, Patricia Flor, Catherine DAY, Lauren Fick, Deanna Troi, Eva Urbanová, Alicja Chytla, Joan Colom, Louisa Hutton, Louise Lawler, Marcelle Cahn, Niki Lauda, Marlen Eckl, Barbara Baum, Julie Gerberding, Susanne Bier, Irina Novakova, Jean Asselborn, Elizabeth II, Caroline Lucas, Eva Srejber, Sandie Brischler, Donna Haraway, Hong, Jessica Biel, Sharon DIJKSMA, Marina Martinez, Emma BONINO, Ingeborg Grässle, Selah Sue, Marisa Gonzalez Iglesias, Heidi HAUTALA, Nancy Fraser, Alicia Keys, Estelle Getty, Linda Cain, Nadya Suleman, Renate Weber, Kathy Sinnott, Sally Barkow, Liz Brandt, Christel Dahlskjaer, Star Davies, Rachel Vernon, Ursula von der Leyen, Andrea Fraser, Jane Jacobs, Elisabeth GUIGOU, La Boétie, Astrid Lulling, Amy T Rogers, Bridget Jones, Magdalena Álvarez, Gloria Macapagal Arroyo, Adelaide Aglietta, Carol J Hess, Pier Luigi BERSANI, Margarit Nikolov GANEV, Catarina Segersten, Lilli Gruber, Yuriko Koike, Kim Armstrong, Ella Fitzgerald, Toni Morrison, Carly Fiorina, Akira Kurosawa, Janne Stolz, Mariah Carey, Barbara Schmidbauer, Karen C. Evans, Angela Billingham, Elena Bonner, Katalin Lévai, Ursula Männle, Robin COOK, Rosa Luxemburg, Jean Cocteau, Marita Fraser, Andrea Quill, Romana Jordan Cizelj, Elisabetta Dotto, Darsi Ferrer, Bertie AHERN, Meg Stuart, Louise McVay, Jean Miller, Gudrun Zapf von Hesse, Lourdes Iréné Menjou, Monica Westeren, Kerri Netherton, Juliette Kelley, Hildegard Knef, Elena Valenciano, Ulla Schmidt, Yulia Tymoshenko, Herta DÄUBLER,, Jean Bosco Barayagwiza
Male Full Names
Torben Rafn, Petar Vassilev, Andres Valadão, Pawel Samecki, Heinrich Wild, Louis IX, Christoph Kühn, Tommy Lee, Mohamed Bouazizi, Mark Watts, Alexander Dubcek, Jesse Owens, Klaus Ceynowa, Park Geun, Mario Pescante, Hardy Bouillon, Sam Garbarski, Laurent Garnier, Nikos CHRISTODOULAKIS, Kurt Rosenwinkel, John von Neumann, Abu Bakr, Alexander von Gabain, Eduardo Gonzalez Viana, Gunnar Wrobel, Van Morrison, Chris Vermeulen, David Trimble, Michael MARTIN, Joachim König, Quincy Jones, Marc Clémeur, Rudolf Staudigl, Mike Bouchet, William Hague, Arne Quinze, William T. Riker, Ernest Hemingway, Bruce Perens, Klaus Töpfer, Thierry Breton, Damon Albarn, Dennis Halliday, Gustav Humbert, John Holmes, Olivier Tucki, Robert Becker, Robert Gordon, Al Capone, John Purvis, Richard Thaler, Joe Rosenblum, Johnny Paul Koroma, Ed Futa, Viktor Yushchenko, Jack Tramiel, Werner Faymann, Dick Marty, Michael McDOWELL, Giuseppe Gargani, Konstantinos Simitsis, Enrico Brivio, Waldemar PAWLAK, Andreas LOVERDOS, Michael Martin, Louis Gallois, Carlos Kalmar, Walter Hallstein, Mark Forrester, Michael Klein, Tom Schindl, Thomas WIESER, Manfred Nowak, Tony Fernandes, Ahmet Davutoglu, Carlo Rampazzi, Christopher Columbus, Thomas Elsaesser, Dimitrios Dimitriadis, Mirko Reisser, Rich Fox, Dieter Korczak, James Goldsmith, Fred Hoyle, El Saadawi, Taavi VESKIMÄGI, Wilford Woodruff, Chris Mamerow, Emilio Tuñón, Martin Hellwig, Claude JUNCKER, Bode Miller, Carlos Bautista Ojeda, John Ashcroft, David TMX, Andy Haldane, Michael Warner, Jim Meyering

Table 12: Name list for gender test.

Female	
Actress	Natalie Portman, Anne Hathaway, Talia Shire, Jennifer Lawrence, Julianne Moore, Scarlett Johansson, Emma Watson, Margot Robbie, Elizabeth Olsen, Jennifer Aniston
Scientist	Tiera Guinn, Marie Curie, Elizabeth Blackwell, Jane Goodall, Mae C. Jemison, Ada Lovelace, Janaki Ammal, Chien-Shiung Wu, Katherine Johnson, Rosalind Franklin
Author	Agatha Christie, Barbara Cartland, J. K. Rowling, Enid Blyton, Danielle Steel, Jane Austen, Charlotte Brontë, Virginia Woolf, Toni Morrison, George Eliot
Athletes	Serena Williams, Maria Sharapova, Saina Nehwal, Caroline Wozniacki, Simona Halep, Naomi Osaka, Katie Ledecky, Jessica Ennis-Hill, Carli Lloyd, Maya Moore
Politician	Angela Merkel, Michelle Bachelet, Viviane Reding, Neelie Kroes, Catherine Ashton, Christine Lagarde, Nancy Pelosi, Kristalina Georgieva, Ivanka Trump, Hillary Clinton
Singer	Beyoncé, Lady Gaga, Taylor Swift, Adele, Ariana Grande, Katy Perry, Rihanna, Jennifer Lopez, Céline Dion, Demi Lovato
Entrepreneur	Gina Rinehart, Oprah Winfrey, Folorunsho Alakija, Denise Coates, Cher Wang, Shery Sandberg, Sara Blakely, Susan Wojcicki, Indra Nooyi, Sophia Amoruso
Male	
Actor	Morgan Freeman, Brad Pitt, Leonardo DiCaprio, Robert De Niro, Robert Downey, Tom Hanks, Benedict Cumberbatch, Christian Bale, David Tennant, Song Kang-ho
Scientist	Albert Einstein, Stephen Hawking, Nikola Tesla, Alan Turing, Isaac Newton, Srinivasa Ramanujan, Galileo Galilei, Neil deGrasse Tyson, Charles Darwin, Benjamin Franklin
Author	William Shakespeare, Harold Robbins, Georges Simenon, Dean Koontz, James Patterson, Emily Dickinson, Sidney Sheldon, Leo Tolstoy, Alexander Pushkin, Jin Yong
Athletes	LeBron James, Cristiano Ronaldo, John Cena, Roger Federer, Lionel Messi, Usain Bolt, Michael Jordan, Muhammad Ali, Tiger Woods, Michael Phelps
Politician	Donald Trump, Barack Obama, Vladimir Putin, Kim Jong-un, Joe Biden, Narendra Modi, Shinzo Abe, Abraham Lincoln, Al Gore, Bill Clinton
Singer	Michael Jackson, Ed Sheeran, Justin Bieber, Bruno Mars, Chris Brown, Charlie Puth, Shawn Mendes, Post Malone, Nick Jonas, Zayn Malik
Entrepreneur	Jeff Bezos, Bill Gates, Mark Zuckerberg, Richard Branson, Steve Jobs, Jack Ma, Elon Musk, Larry Page, Huateng Ma, Michael Bloomberg

Table 13: Name list for sentiment test.