Universität Zürich[UZH]

Master's Thesis
Master of Science
in Department of Informatics

Reducing Gender Bias in Neural Machine Translation with FUDGE

Author: Tianshuai Lu
Student ID: 20-743-134

Supervisor: Prof. Dr. Martin Volk
Dr. Annette Rios
Noëmi Aepli

Department of Computational Linguistics

Submission Date: 27.04.2023

Abstract

Gender bias appears in many neural machine translation models and commercial translation software. The problem is well known and efforts to reduce such discriminatory tendencies are underway. But gender bias is still not fully solved. This work utilizes a controlled text generation method, Future Discriminators for Generation (FUDGE), to reduce the so-called *Speaking As* gender bias emerging when translating from English to a language that openly marks the gender of the speaker. The model is evaluated with BLEU and MuST-SHE, a novel gender translation evaluation method. The results demonstrate improvements in the translation accuracy of the feminine terms.

Zusammenfassung

Geschlechterspezifische Tendenzen treten in vielen neuronalen maschinellen Übersetzungsmodellen und kommerziellen Übersetzungssoftware auf. Das Problem ist bekannt und Anstrengungen, solche diskriminierenden Tendenzen zu reduzieren sind im Gange, allerdings soweit ohne durchschlagenden Erfolg. Für diese Arbeit nutzt eine kontrollierte Textgenerierungsmethode namens Future Discriminators for Generation (FUDGE) eingesetzt, um die bei der Übersetzung von Englisch in eine Sprache, die das Geschlecht des Sprechers offen markiert, auftretende sogenannten *Speaking As* Geschlechterdiskriminierung zu reduzieren. Das Modell wird mit BLEU und MuST-SHE, einer neuartigen Methode zur Evaluation geschlechtsspezifischer Übersetzungen, bewertet. Die Ergebnisse zeigen Verbesserungen in der Übersetzungsgenauigkeit von femininem Genus auf.

Acknowledgement

First and foremost I would like to express my gratitude to Prof. Dr. Martin Volk for agreeing to supervise my thesis and giving me the freedom to choose the topic that I enjoy doing.

And I would like to extend my sincere appreciation to Dr. Annette Rios and Noëmi Aepli for their invaluable support and guidance throughout the project. From helping me find appropriate data sets to leading me in the right direction of the project, every bit of help is vital to the success of this project.

A special thanks to Beatrice Savoldi for answering my questions and providing the correct extension files of the MuST-SHE benchmark, without which the evaluation wouldn't have been conducted smoothly.

Last but not least, I would like to thank all my friends and family, especially my parents, for their unwavering support.

Contents

List of Figures

List of Tables

List of Acronyms

AGI	Artificial General Intelligence
AI	Artificial Intelligence
CCLM	Class-Conditional Language Model
FUDGE	Future Discriminators for Generation
LLM	Large Language Model
LM	Language Model
LSTM	Long short-term memory
MT	Machine Translation
NLP	Natural Langauge Processing
NMT	Neural Machine Translation
POS	Parts-Of-Speech

1 Introduction

1.1 Motivation

When we talk about gender bias in Neural Machine Translation (NMT), the first issue that comes to mind is stereotyping, e.g. associating the profession doctor with the male pronoun and a nurse with the female pronoun. This example is indeed a gender bias, but as Hardmeier et al. (2021) pointed out, there are different kinds of gender bias in NMT and it is necessary to identify what is deemed as harmful behavior, how, and to whom (Savoldi et al., 2021).

In a broader sense, gender bias is found in many Natural Language Processing (NLP) systems, and has raised much concern about gender inequality and the danger of reinforcing damaging stereotypes in downstream applications. With the rapid development of Large Language Models (LLMs), NLP has gained much more attention across the general public. While people are fascinated by its capabilities and what we could achieve with such powerful systems, some researchers recently put out an open letter[1] to call on the labs and research institutes across the world to pause the training of AI systems that are potentially better than GPT-4 (OpenAI, 2023), which is released on March 14, 2023, and suggesting that "powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable". We will not discuss whether it is feasible to seize the development of better LLMs for six months, the concern in the letter is valid and will be a vital topic for our future research in NLP or Artificial General Intelligence (AGI) in general. And gender bias is certainly one of the AI safety problems we need to face and tackle.

Current research on mitigating gender bias in MT usually focuses on gender stereotypes (Stanovsky et al., 2019), translation errors due to speaker gender (Vanmassenhove et al., 2018) or pronoun translation (Loáiciga et al., 2017; Jwalapuram et al., 2020). Furthermore, the proposed methods are usually only evaluated on BLEU (Papineni et al., 2002). However, BLEU evaluates the overall translation fluency and

[1] `https://futureoflife.org/open-letter/pause-giant-ai-experiments/`

is rather insensitive to specific linguistic phenomena that only affect a few words (Sennrich, 2017).

In this work, I apply a controlled text generation method, Future Discriminators for Generation (FUDGE) (Yang and Klein, 2021), to mitigate the gender bias that arises when translating from English, a language that marks gender only on pronouns, to Italian, a language that openly marks the gender of the speaker in specific contexts. FUDGE has demonstrated its capabilities on many controlled text generation tasks, e.g. poetry couplet completion, topic-controlled language generation, and machine translation formality change. With this work, I further explore its performance on a gender-controlled machine translation task.

Furthermore, instead of solely relying on BLEU (Papineni et al., 2002) as an evaluation metric, the models were also evaluated on a novel gender translation challenge set, MuST-SHE (Savoldi et al., 2022), built on manually annotated test sets specifically designed for its purpose. And FUDGE demonstrates improvements in several feminine gender terms' translation accuracy. The code and documentation of this project are available on GitHub[2].

A special note on this project: this work is part of the EASIER project, a Horizon 2020 project that aims to design, develop, and validate a complete multilingual machine translation system that will act as a framework for barrier-free communication among deaf and hearing individuals, as well as provide a platform to support sign language content creation[3].

1.2 Thesis Structure

In the following chapters, we start with reviewing some relevant work in NLP (Chapter 2) and build the context with the help of some sociolinguistic literature (Chapter 3). Next, we introduce the FUDGE method and how we are going to use it to mitigate gender bias (Chapter 4). Then the reader will get more insights into the data (Chapter 5) and the experiment settings (Chapter 6). Afterward, we will examine the results (Chapter 7) and analyze them (Chapter 8). Lastly, we wrap up with a conclusion (Chapter 9).

[2] `https://github.com/tianshuailu/debias_FUDGE`

[3] For more details about the project, please refer to `https://www.project-easier.eu/`

2 Related Work

2.1 Controlled Text Generation

A number of research works concentrate on fine-tuning a pre-trained model for a desired attribute. Ficler and Goldberg (2017) proposed a framework for neural natural language generation (NNLG) controlling different stylistic aspects of the generated text. And the method results in a class-conditional language model (CCLM), but it is usually difficult to separate the desired attribute from the generation model, which means the model is usually just suitable for one task and needs retraining for another attribute of interest. Keskar et al. (2019) mitigates this issue by proposing a Conditional Transformer Language (CTRL) model that was trained to condition on many factors, e.g. style, content, and task-specific behavior. This method, however, is quite expensive. Furthermore, Krause et al. (2021) suggests using discriminators to guide the decoding of LMs.

Another line of work achieves the goal of controlling attributes by backpropagating the gradients. Dathathri et al. (2020) proposed Plug and Play Language Model (PPLM) for controllable language generation, also utilizing attribute classifiers to guide the decoding, but with a forward and a backward gradient path to push the LM's hidden activations and thus control the generation.

2.2 Mitigating Gender Bias

Gender Tagging

A common method to mitigate gender bias is to attach gender tags as proposed by Vanmassenhove et al. (2018). In this case, gender information is integrated into the Neural Machine Translation (NMT) systems via a tag on the source. This approach achieves improvement for multiple language pairs.

Data Augmentation

Given the original biased data set, Zhao et al. (2018) proposed to construct an additional training corpus where all male entities are swapped for female entities and vice-versa. The goal of the augmentation is to mitigate the bias by training the model on gender-balanced data sets. And Zhao et al. (2018) claimed that data augmentation reduces gender bias and works even better in combination with removing biased resources.

Benchmarks

Zhao et al. (2018) introduced a benchmark, WinoBias, to measure gender bias in coreference resolution with entities corresponding to people referred to by their occupation. Another benchmark, WinoGender (Rudinger et al., 2018), is a Winograd schema-style (Levesque et al., 2012) set of minimal pair sentences that differ only by pronoun gender.

Based on WinoBias and WinoGender, Stanovsky et al. (2019) composed a coreference resolution English corpus that contains sentences in which the subjects are in non-stereotypical gender roles. It is a standard test set to evaluate gender stereotyping in MT.

3 Sociolinguistics Background

Before diving into how to mitigate gender bias in Neural Machine Translation, let us take a detour to some sociolinguistic literature to have a general impression of what gender bias is, where it comes from, and what the different types of gender bias in Machine Translation are.

3.1 Gender and Language

The research on gender differences in language is multidisciplinary and may involve social psychology, anthropology, linguistics, and sociolinguistics. From the work of earlier anthropologists, we have a glimpse of many phonological, morphological, and lexical differences between female and male speakers (Coates, 2004). In social psychology, researchers reveal the cause and the consequences of gender differences in language (Stahlberg et al., 2007), while a linguist might be more interested in how gender difference is reflected in the grammatical structure or lexicon of a language.

3.1.1 Gender Representation in Language

Gender, used to describe socially constructed categories, is also reflected in the grammatical structures of almost all languages. The work by Stahlberg et al. (2007) explores how gender is expressed in different languages and groups languages into three categories in terms of gender representation: grammatical gender languages, neutral gender languages, and genderless languages.

Grammatical Gender Languages

In grammatical gender languages (e.g. Italian, Spanish), every noun is assigned a feminine or masculine (or possibly neuter) gender. And other Parts-Of-Speech, e.g. verbs, articles, adjectives, and pronouns are also marked with the gender of the nouns.

Finnish	Hän on hyvä ystävä.
English	She/He is a good friend.
Italian	Lei/Lui è una/un buona/buon amica/amico.

Table 1: An example of two sentences written in each one of the three types of languages, namely grammatical gender languages (Italian), natural gender languages (English), and genderless languages (Finnish).

Natural Gender Languages

In natural gender languages, like English, most personal nouns (e.g. doctor, nurse, student, and teacher) can be used to refer to both females and males. There is no grammatical marking of gender, hence verbs, articles, and adjectives are usually the same for females and males. Expression of gender is less frequent and easier to avoid in natural gender languages than grammatical gender languages (Stahlberg et al., 2007).

Genderless Languages

Genderless languages (e.g. Finnish, Turkish), as defined by Stahlberg et al. (2007), have neither grammatical gender in the noun system nor gender-differential personal pronouns. And gender is usually only expressed in lexical pairs (e.g. in Finnish, *nainen*/woman vs. *mies*/man). In addition, verbs, articles, and adjectives are also not marked with gender.

In summary, all three types of languages have lexical expressions of gender. Apart from that, pronominal forms are used to express gender in natural gender languages and grammatical gender languages. And only grammatical languages possess gender-marked nouns and other Parts-Of-Speech that grammatically agree with them. In Table 1, two sentences are written in each one of these three types of languages. For the English sentence, the only difference is the pronoun *He* or *She*, while in the Italian version the adjectives, determiners, and personal nouns also change with the pronoun. But in the Finnish example, we could not determine the gender of the person being referred to just by looking at the sentence.

A note on the aforementioned classification: it is not a rigorous linguistic categorization, but a theory suggested by social psychologists Stahlberg et al. (2007) and fits in a specific research context that may help with the discussion of gender bias in NMT.

3.2 Gender and Bias

A systematic classification of gender bias in MT does not seem to exist but would be necessary when tackling specific gender bias with specific data sets and methods. In this project, I consider two classification methods: one accounting for the different sources that can lead to machine bias (Friedman and Nissenbaum, 1996), and another one approaching it from a linguistic standpoint, proposed by Dinan et al. (2020).

3.2.1 Bias Categories according to source

Friedman and Nissenbaum (1996) proposed three categories of bias in computer systems in general, pre-existing, technical, and emergent. The pre-existing bias is rooted in our institutions, practices, and attitudes; the technical bias is caused by technical constraints and decisions; the emergent bias arises from the interaction between systems and users (Savoldi et al., 2021).

Pre-existing Bias

MT models are known to reflect gender disparities present in the data (Savoldi et al., 2021). The data sets that were used to train the MT models are generated from our everyday life, which inevitably carries a social, historical, and cultural context, including bias, into the models.

Technical Bias

Data sets are one of the main sources where gender bias comes from, but as Geirhos et al. (2020) pointed out, data alone rarely constrains a model sufficiently, and data cannot replace making assumptions. As demonstrated in multiple works, the model architecture, the choices of the parameters, and algorithms all could potentially influence the models' behaviour (Vanmassenhove et al., 2019; Escolano et al., 2021; Roberts et al., 2020).

Emergent Bias

While the pre-existing bias and the technical bias can be identified before or during the development of MT models, emergent bias only happens after using the models, usually as a result of a change in societal knowledge or cultural values. Google Translate was launched in 2006, but it wasn't until 2018 that it supported the translations of feminine and masculine forms[1]. With the awareness of non-binary

[1] `https://ai.googleblog.com/2018/12/providing-gender-specific-translations.html`

gender groups and gender-inclusive language, more improvements certainly need to be done for the MT models.

3.2.2 ABOUT, AS, TO Framework

To tackle the challenge of gender bias in NMT or NLP in general, we need a fine-grained classification of gender bias derived from NLP systems. Dinan et al. (2020) proposed a multi-dimensional framework (ABOUT, AS, TO) for measuring gender bias in language and NLP models.

Speaking About: Gender of the Topic.

This type of bias is probably the most well-known in NLP, e.g. the notorious example that assumes a doctor is male and a nurse is female. This dimension, *Speaking About*, is often embedded in the text composed of third-person sentences and is usually caused by the under-representation of certain social groups (Savoldi et al., 2021).

Speaking To: Gender of the Addressee.

We expect a person to adjust their speech depending on the gender of whom they are speaking to (Eckert and Rickford, 2001). This dimension, *Speaking To*, refers to the gender bias that emerged around the gender of the addressee.

Speaking As: Gender of the Speaker.

When we use the pronoun *I*, we bring to that use of *I* a sense of being either a woman or a man (Coates, 2004). Coates (2004) extensively summarized the gender differences in languages with diverse examples demonstrating that women and men have different linguistic behavior.

In the experiments, the last dimension, *Speaking As*, will be the focus of this work. The training data sets, test sets, and evaluations are all selected according to this specific type of gender bias.

3.3 Remarks

In light of the above multidisciplinary background, we formed the context of this work, which will focus on mitigating the *Speaking As* type of gender bias that occurs when translating from a natural gender language to a grammatical gender language i.e. from English to Italian, under the assumption that the bias pre-exists in the data sets.

4 FUDGE

4.1 Motivation

Controlled text generation is the task of generating text conditioned on an additional desirable attribute a which is not already built into the trained model $\mathcal{G}$ (Yang and Klein, 2021). For example, the model $\mathcal{G}$ is a translation model from English to Italian, and we would like the output to possess some attributes that it does not already have, e.g. we would like the Italian translation to be more formal, which is not optimized during training.

Unfortunately, it is usually nontrivial to retrain the model $\mathcal{G}$ to condition it on the new attribute a, and if new attributes come up, the training process will need to be repeated. Another possible solution suggested by Keskar et al. (2019) is a Conditional Transformer Language (CTRL) model that was trained to condition on many factors. But it is only limited to the pre-defined factors, which still does not solve the problem of retraining the model for a new attribute.

Considering these limitations, Yang and Klein (2021) proposed Future Discriminators for Generation (FUDGE), a flexible and modular way of conditioning the generative model $\mathcal{G}$ on the desired attribute a that only requires access to the output probabilities of $\mathcal{G}$. FUDGE achieves this by training a binary classifier that predicts whether the attribute a is satisfied in the complete sequence based on the generated sequence for each step.

4.2 Future Discriminators for Generation

In a controlled text generation task, assume an autoregressive model $\mathcal{G}$ that samples from a distribution $P(X)$:

$$P(X) = \prod_{i=1}^{n} P(x_i \mid x_{1:i-1})$$

To condition on the desired attribute a, it requires to model $P(X \mid a)$:

$$P(X \mid a) = \prod_{i=1}^{n} P(x_i \mid x_{1:i-1}, a)$$

instead of modeling a class-conditional language model, FUDGE utilizes a Bayesian factorization:

$$P(x_i \mid x_{1:i-1}, a) \propto P(a \mid x_{1:i}) P(x_i \mid x_{1:i-1})$$

The second term $P(x_i \mid x_{1:i-1})$ is the prediction of $\mathcal{G}$, which makes it suffice to model the first term $P(a \mid x_{1:i})$ with a binary classifier $\mathcal{B}$. (Yang and Klein, 2021)

4.2.1 Advantages and Limitations

In the work by Yang and Klein (2021), they mentioned a few potential advantages of FUDGE compared to other methods e.g. directly fine-tuning a class-conditional language model (CCLM) on $\mathcal{G}$:

1. FUDGE only requires access to the output logits of $\mathcal{G}$, instead of $\mathcal{G}$ itself.
2. When a better model is available, $\mathcal{G}$ can be replaced as long as it shares the same tokenization as $\mathcal{G}$.
3. When a task requires multiple conditionally independent attributes, FUDGE can easily combine the classifiers for each attribute by summing their output log probabilities.

One drawback is that FUDGE can not guarantee that all the outputs meet the desired attribute a due to the approximation when modeling $P(a \mid x_{1:i})$ and only considers the top 200 tokens for computational efficiency.

4.2.2 Example Applications

In the original work by (Yang and Klein, 2021), FUDGE was tested on three different controlled text generation tasks: poetry couplet completion, topic-controlled language generation, and machine translation formality change.

Poetry Couplet Completion

Given the first line of an English iambic pentameter couplet, the task is to generate a second line that satisfies the iambic pentameter, rhymes with the first line, and ends a sentence. For this task, Yang and Klein (2021) built three classifiers for

the three components of the attribute a: meter, rhyme, and sentence-ending, then combined them at test time. Compared to the baselines, FUDGE maintains a fluent generation and increases diversity.

Topic-Controlled Language Generation

In this example, Yang and Klein (2021) experimented on English topic control language generation. Given a topic and a generic prefix, the goal is to generate three 80-token samples. There are seven topics and 20 prefixes. The desired attribute a is to be on-topic for a given topic. Given a word and the prefix, the classifier predicts whether the word will appear in the future. FUDGE outperforms the baselines in both topic relevance and fluency.

Machine Translation Formality Change

Yang and Klein (2021) describes Machine Translation Formality Change as a somewhat more challenging task, which translates a conversational Spanish sentence into a formal English sentence. The desired attribute a is formality without sacrificing the meaning of the source sentence. The binary classifier classifies whether the text starting with a certain prefix is written in a formal style. In evaluation, while FUDGE and $\mathcal{G}$ achieve similar BLEU scores, FUDGE has a higher formality score.

For more details about these three tasks, please refer to the paper (Yang and Klein, 2021). Of all three applications, Machine Translation Formality Change shares the most similarities with gender-controlled machine translation.

4.3 Gender-Controlled Machine Translation

Given an English sentence, the goal is to generate a semantically correct Italian translation with proper grammatical gender agreement according to the speaker's gender. Since the gender of the speaker is unmarked in the English sentence, I experimented on feminine and masculine genders separately, meaning that one model will only combine with the feminine classifier and another one only with the masculine classifier. In essence, the experiments combine the underlying English–Italian translation model with a binary classifier that predicts whether the completed sequence will have feminine or masculine forms according to the speaker's gender.

To see if gender tags improve FUDGE's performance, I tested two underlying translation models $\mathcal{G}$ and $\mathcal{G}_t$. The only difference between $\mathcal{G}$ and $\mathcal{G}_t$ is that $\mathcal{G}_t$ utilized gender tags. $\mathcal{G}$ and $\mathcal{G}_t$ are also the baselines of the experiments in section 6.

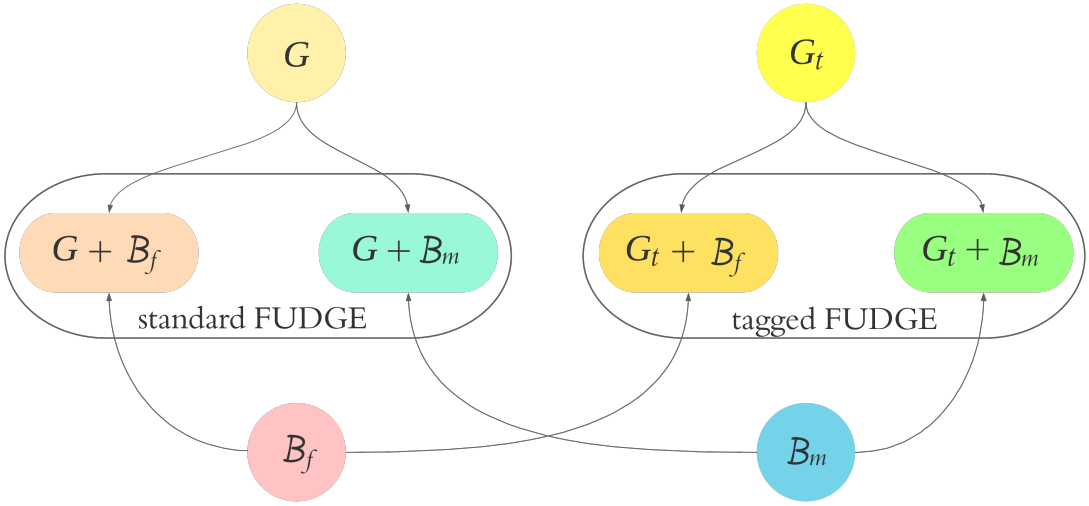


Figure 1: Illustration of four combinations between the underlying translation models $\mathcal{G}$ (translation model trained on original data sets), $\mathcal{G}_t$ (translation model trained on tagged data sets) and two classifiers $\mathcal{B}_f$ (feminine), $\mathcal{B}_m$ (masculine).

In order to translate with correct grammatical gender agreement, FUDGE requires two desired attributes a, feminine and masculine, hence two classifiers $\mathcal{B}_f$ and $\mathcal{B}_m$. *Each* of them is combined with the two underlying translation models $\mathcal{G}$ and $\mathcal{G}_t$, resulting in four combinations, as illustrated in Figure 1.

FUDGE relies on only one hyperparameter, λ, to control the weight of the classifiers ($\mathcal{B}_f$ and $\mathcal{B}_m$) versus G's prediction. In the experiments, the translation models were combined with the classifiers on different values of λ, ranging from 1 to 5.

4.4 Standard Baseline and Tagged Baseline

One baseline was fine-tuned on an English–Italian parallel data set as a standard baseline, whereas the tagged baseline was fine-tuned on an identical data set but with a gender tag attached to each English source sentence, tagged baseline. The tagged baseline was built with the method from Vanmassenhove et al. (2018).

Together with the release of Europarl-Speaker-Information corpora (Vanmassenhove and Hardmeier, 2018), Vanmassenhove et al. (2018) proposed to incorporate gender information into neural machine translation (NMT) systems. Similar to the work by Sennrich et al. (2016b) on controlling politeness by marking the source side

of the training data with a sentence level feature tag "informal" or "polite", Vanmassenhove et al. (2018) augmented the source side sentences with the gender tag "FEMALE" or "MALE", as illustrated in the example (3.1):

(4.1) *"FEMALE Madam President, as a..."*

As an advantage of FUDGE, it only needs access to the output logits of the generator model, which means the baseline models can be used as the generator models and can be directly combined with the classifiers without additional fine-tuning or modification. This allows me to directly use $\mathcal{G}$ and $\mathcal{G}_t$ as baselines.

5 Data

5.1 Europarl-Speaker-Information

Europarl (Koehn, 2005) is a corpus of parallel sentences in eleven languages from the proceedings of the European Parliament. Vanmassenhove and Hardmeier (2018) tagged them with speaker information (name, gender, age, date of birth, euroID, and date of the session) based on monolingual Europarl (Koehn, 2005) source files which contain speaker names on the paragraph level.

I chose Europarl-Speaker-Information (Vanmassenhove and Hardmeier, 2018) because it is the largest dataset with parallel sentences tagged with the gender of the speaker. In addition, Europarl (Koehn, 2005) consists of parliament proceedings, hence it might contain more first-person sentences, which makes it suitable for the kind of gender bias the experiments focus to reduce, i.e., *Speaking As*. To verify this assumption, I selected two crawled corpora and two news corpora from OPUS[1] and compared their number of sentences that contain *I* or *we* on the English side with Europarl-Speaker-Information, as shown in Table 2. The two crawled corpora are ParaCrawl (Bañón et al., 2020) and CCAligned (El-Kishky et al., 2020). The two news corpora are News-Commentary (Tiedemann, 2012) and GlobalVoices (Tiedemann, 2012). Numbers in Table 2 were obtained on the English–Italian parts of the corpora.

I used Europarl-Speaker-Information (Vanmassenhove and Hardmeier, 2018) to train the baseline and tagged baseline and test the FUDGE (Yang and Klein, 2021) method. The English–Italian part of the corpus contains 1.29 million sentences, which includes 870,000 sentences by male speakers and 420,000 sentences by female speakers. Hence the ratio of the two gender groups is 2.08/1. Before training the tagged baseline, I attached the gender tag to each English source sentence, indicating the gender of the speaker, as shown in the two examples:

[1] `https://opus.nlpl.eu/`

corpus	**first-person sentences**	**corpus size**	**percentage**
News Commentary	4,422	78,348	5.64%
ParaCrawl	7,392,692	96,977,931	7.62%
CCAligned	1,037,340	13,186,311	7.87%
GlobalVoices	13,844	122,074	11.34%
Europarl-Speaker-Information	576,402	1,295,563	**44.49%**

Table 2: A comparison of the number of first-person sentences in the selected corpora by counting the number of sentences that contain *I* or *we* on the English sentence side. Europarl-Speaker-Information(Vanmassenhove and Hardmeier, 2018) has the highest percentage of first-person sentences.

	standard FUDGE	**tagged FUDGE**
baseline training set	1.2 million	1.2 million
validation set	2000	2000
feminine test set	1300	1300
masculine test set	2700	2700

Table 3: The number of sentences for each data set. The sizes of data sets for standard FUDGE and tagged FUDGE are the same because I used the same sentence pairs for FUDGE and tagged FUDGE.

(5.1) *[FEMALE] I want to refer to two areas that I know well.*

(5.2) *[MALE] I believe it is necessary to return to ...*

To individually evaluate the feminine and masculine classifiers, a sample of 4000 sentences was randomly selected and divided according to gender, yielding 1300 sentences by female speakers and 2700 sentences by male speakers, which share the same gender ratio of the corpora. To ensure comparability of the results, standard FUDGE and tagged FUDGE utilized identical sentence pairs, with the only difference being the inclusion of gender tags in tagged FUDGE's data sets. Table 3 provides more details about the data sets.

	unprocessed	processed	filtered
feminine	24,700	283,500	45,800
masculine	54,700	713,000	248,900

Table 4: The line counts of ParlaMint 2.1 feminine and masculine data sets. The first two columns display the line counts before and after splitting the utterances into one sentence per line. The last column displays the line counts after filtering out the sentences that do not contain adjectives or participles.

	unprocessed	processed	filtered
feminine	318	26	38
masculine	351	25	34

Table 5: The average length in tokens of ParlaMint 2.1 feminine and masculine data sets. The length is measured by tokens per sentence. The first two columns display the lengths from before and after splitting the utterances into one sentence per line. The last column displays the lengths after filtering out the sentences that do not contain adjectives or participles.

5.2 ParlaMint 2.1

ParlaMint 2.1 (Erjavec et al., 2021) is a multilingual set of 17 corpora containing parliamentary debates mostly starting in 2015 and extending to mid-2020. The corpora have extensive metadata, including aspects of the parliament and the speakers (name, gender, and more). The gender of the speaker is more relevant to the experiments. Besides, the speeches also contain marked-up transcriber comments, such as gaps in the transcription, interruptions, applause, etc.

I selected ParlaMint 2.1 to train the classifiers for two reasons. Firstly, it has an Italian corpus with a sufficient amount of sentences and is tagged with the speaker's gender. In addition, similar to Europarl-Speaker-Information, it is also parliamentary proceedings.

The corpus is divided into feminine and masculine data sets, making it suitable to train the feminine and masculine classifiers separately. Each line contains the whole speech of the speaker, so I split the utterances into one sentence per line with

the sentencizer[2] from spaCy (Honnibal et al., 2020). Furthermore, I removed the transcriber comments, since they are irrelevant to the gender of the speaker. To provide a general impression, the line counts and the average lengths of feminine and masculine data sets before and after processing are displayed in the first two columns of Table 4 and Table 5. The length is measured by tokens per sentence.

In Italian, the adjectives and participles are marked with the gender of the speaker. As shown in Table 5, in the full data set, the utterances where the gender of the speaker is marked are relatively sparse. Hence it is possible that using the full data set does not give the classifier a strong enough signal to learn. While almost all of the sentences in the filtered data set contain gender-marked words, hence it should be more suitable to train the classifier. The accuracy of classifiers trained on these two data sets is shown in section 7.2. The sentences that contain adjectives and participles were selected with the Italian pipeline[3] and the Morphologizer[4] from Spacy (Honnibal et al., 2020). The line counts and the average length are shown in the third columns of Table 4 and Table 5.

5.3 MuST-SHE v1.2

MuST-SHE (Bentivogli et al., 2020) is a multilingual benchmark allowing for a fine-grained analysis of gender bias in Machine Translation and Speech Translation and is a subset of the TED-based MuST-C corpus (Di Gangi et al., 2019). The dataset comprises audio, transcript, and translation triplets annotated with the gender of the speaker. Only the English transcript and Italian translation pairs were used during the experiments. Each pair requires translating at least one English gender-neutral word into the corresponding masculine/feminine target word(s). (Bentivogli et al., 2020)

WinoMT (Stanovsky et al., 2019) represents the standard corpus to evaluate gender bias in MT within an English-to-grammatical gender language scenario (Savoldi et al., 2022). While WinoMT might be suitable for detecting gender stereotypes, it consists of synthetic sentences with the same structure (e.g. *The CEO helped the nurse because he needed help.*) and doesn't include first-person sentences. On the other hand, MuST-SHE v1.2 is built on the utterances from TED talks and contains 656 first-person sentences out of 1073. Hence, I selected MuST-SHE v1.2 over WinoMT for FUDGE's evaluation.

[2] `https://spacy.io/api/sentencizer`

[3] `https://spacy.io/models/it`

[4] `https://spacy.io/api/morphologizer`

	Quantity
She	554
He	511
They	10
Multi-She	4
Multi-He	4
Multi-Mix	12
Sum	1095

Table 6: The distribution of gender tags in the en-it MuST-SHE benchmark. *They* means the speaker is referred to with gender-neutral linguistic gender forms. *Multi-She/He* means two or more speakers who are all referred to with feminine/masculine linguistic gender forms. *Multi-Mix* means two or more speakers who are referred to with different linguistic gender forms.

MuST-SHE v1.2 categorizes the gender into six categories: *She, He, They, Multi-She, Multi-He* and *Multi-Mix*. *They* means the speaker is referred to with gender-neutral linguistic gender forms. *Multi-She/He* means two or more speakers who are all referred to with feminine/masculine linguistic gender forms. *Multi-Mix* means two or more speakers who are referred to with different linguistic gender forms. Since only feminine and masculine linguistic gender forms are considered in the experiments, I only kept the sentences of these four categories: *She, He, Multi-She* and *Multi-He* and merged the sentences labeled with *Multi-She* and *She*, *Multi-He* and *He*. The distribution of these six categories is shown in Table 6.

MuST-SHE v1.2 (Savoldi et al., 2022) adds extensions consisting of two manually created linguistic annotation layers, which enrich MuST-SHE (Bentivogli et al., 2020) with information about Parts-Of-Speech and gender agreement chains.

Parts-Of-Speech

Savoldi et al. (2022) annotate each gender-marked word with POS information that is differentiated among six categories: 1) articles, 2) pronouns, 3) nouns, 4) verbs. Additionally, adjectives are divided into 5) *limiting adjectives* or adj-determiner, adjectives with minor semantic import that determine e.g. possession, quantity, space (my, some, this); 6) *descriptive adjectives* or adj-descriptive that convey attributes and qualities, e.g. glad, exhausted. This classification of adjectives is from the work by Schachter and Shopen (2007).

	Quantity
POS(total)	2026
Art	413
Pronoun	48
Adj-det	149
Adj-des	448
Noun	346
Verb	622
AGR-CHAINS	421

Table 7: The distribution of Parts-Of-Speechb(POS) and the number of gender agreement chains (AGR-CHAINS) in the en-it MuST-SHE benchmark. Adj-det denotes the determiner adjective and Adj-des denotes the descriptive adjective.

For the purpose of word-level evaluation, Savoldi et al. (2022) only consider the words that are subject to form variations due to gender morphological inflections and categorize them into the six aforementioned types:

1. *Art* includes articles and prepositional articles. Unlike simple prepositions (e.g. of, in) that are invariable with gender, prepositional articles combine a preposition and an article and are variable with gender, as shown in example (5.3) [5]:

   (5.3) *delle* *vecchie* *signore*
   of the old ladies

2. *Noun* only includes human-referring nouns

3. *Verb* includes all the verbs that are inflected for gender agreement with the subject.

4. *Adjective* are further divided into adj-determiner (5.4) and adj-descriptive (5.5):

   (5.4) *C'erano* *parecchi* *ragazzi*
   There were several guys

[5]The English–Italian examples from (5.3) to (5.11) and the following categorical information are from the Annotation Guidelines of MuST-SHE v1.2 Extensions `https://ict.fbk.eu/must-she/`.

(5.5) *la* *signora* *vecchissima*
the woman very old
'the very old woman'

5. *Pronoun* is used to annotate pronouns that are also marked with gender.

Agreement

Savoldi et al. (2022) also enrich MuST-SHE (Bentivogli et al., 2020) with grammatical gender agreement (Corbett, 2006), which requires that related words match the same gender form, as in the case of phrases, i.e. groups of words that constitute a single linguistic unit. They identify and annotate as agreement chains gender-marked words that constitute a phrase, such as a noun plus its modifiers, and verb phrases for compound tenses. (Savoldi et al., 2022)

Agreement-level annotation concerns the words that constitute a phrase, to examine whether they agree in gender inflection. Consider two Italian translations of the English phrase *A goof friend*, (5.6) is a correct translation with all words concording to a feminine agreement.

(5.6) *Una* *brava* *amica*
A good friend

While in (5.7), the Italian phrase is ungrammatical, since *Un* is in masculine form while *brava* and *amica* are in feminine form.

(5.7) **Un* *brava* *amica*
A good friend

For the purpose of the gender agreement level evaluation, Savoldi et al. (2022) consider three types of phrases:

1. Noun phrases, phrases that have a noun as their head. Within noun phrases, determiners and adjectives are also considered as shown in example (5.8)

   (5.8) *L'* *altro* *volontario*
   The other volunteer

2. Prepositional phrases, phrases that have prepositions as their head, which usually have the structure **Prepositional articles + Noun**, as seen in example (5.9)

(5.9) *coi* *cuochi*
with the cooks

3. Verb phrases, phrases that consist of a main verb alone, or a main verb plus any modal and/or auxiliary verbs with a predicative function. Savoldi et al. (2022) further consider verbal (5.10) and nominal (5.11) predicates two cases.

(5.10) *sono* *stato* *chiamato* *per* *l'incontro*
I've been called for the meeting

(5.11) *sono* *diventato* *un* *musicista*
I became a musician

5.4 Overview

Finally, as an overview of how each data set is utilized, Table 8 presents the purpose of the three data sets. Europarl-Speaker-Information (Vanmassenhove and Hardmeier, 2018) English–Italian parallel data sets were used to train and calculate the BLEU score of the underlying translation models $\mathcal{G}$ and $\mathcal{G}_t$. $\mathcal{G}$ and $\mathcal{G}_t$ were trained on the same sentence pairs, except that $\mathcal{G}_t$'s training set contains gender tags on the English source side.

ParlaMint 2.1 (Erjavec et al., 2021) Italian monolingual data sets were used to train and test the feminine and masculine classifiers $\mathcal{B}_f$ and $\mathcal{B}_m$. $\mathcal{B}_f$ and $\mathcal{B}_m$ were trained on the same monolingual sentences, except that $\mathcal{B}_f$ treats the feminine sentences as positive and $\mathcal{B}_m$ the masculine sentences.

MuST-SHE v1.2 (Savoldi et al., 2022) English–Italian parallel data sets were only used to evaluate the standard and tagged FUDGE on the word level and gender agreement level.

	Type	**Usage**
Europarl-Speaker-Information	en-it parallel	training and testing $\mathcal{G}$ and $\mathcal{G}_t$
ParlaMint 2.1	it monolingual	training and testing $\mathcal{B}_f$ and $\mathcal{B}_m$
MuST-SHE v1.2	en-it parallel	evaluation

Table 8: An overview of the language type and the usage of the corpora. Europarl-Speaker-Information (Vanmassenhove and Hardmeier, 2018) English–Italian parallel data sets were used to train and test the underlying translation models $\mathcal{G}$ and $\mathcal{G}_t$. ParlaMint 2.1 (Erjavec et al., 2021) Italian monolingual data sets were used to train and test the feminine and masculine classifiers $\mathcal{B}_f$ and $\mathcal{B}_m$. MuST-SHE v1.2 (Savoldi et al., 2022) English–Italian parallel data sets were used to evaluate FUDGE and tagged FUDGE.

6 Experimental Setup

6.1 Models

6.1.1 Underlying English–Italian translation models

The two underlying translation models $\mathcal{G}$ and $\mathcal{G}_t$ are the results of fine-tuning mT5 (Xue et al., 2021). mT5 is a multilingual pre-trained text-to-text transformer, a multilingual variant of T5 (Raffel et al., 2020), and was pre-trained on a Common Crawl-based dataset mC4 (Xue et al., 2021) covering 101 languages. To cover these languages, the vocabulary size of mT5 is 250,000 words. Considering the experiments only use English and Italian, this amount of vocabulary is redundant, which makes it necessary to prune mT5 with a smaller vocabulary.

The first step of the experiments is trimming the model mT5-small (Xue et al., 2021) from Hugging Face[1] with a smaller vocabulary of 25,000 English and Italian subword entries. To get these subword entries, I applied byte-pair-encoding (BPE) (Sennrich et al., 2016a) to the English and Italian sentences from Europarl-Speaker-Information (Vanmassenhove and Hardmeier, 2018) and ranked the subword units by frequency and selected the top 25,000 items as the vocabulary to trim mT5.

The trimmed mT5 (Xue et al., 2021) was fine-tuned on Europarl-Speaker-Information (Vanmassenhove and Hardmeier, 2018) English–Italian data set with the example scripts from Hugging Face Transformers (Wolf et al., 2020), yielding the underlying translation models $\mathcal{G}$ and $\mathcal{G}_t$. The scripts were adapted by adding an early stopping callback[2] and the early stopping patience was set to 10.

$\mathcal{G}$ and $\mathcal{G}_t$ were trained on the same Europarl-Speaker-Information (Vanmassenhove and Hardmeier, 2018) data sets, except that $\mathcal{G}_t$'s English source side has a gender tag, either "[FEMALE]" or "[MALE]", attached to each line. But $\mathcal{G}$ and $\mathcal{G}_t$ shared the same model architecture and training hyperparameters. Besides, $\mathcal{G}$ and $\mathcal{G}_t$ will

[1]The link of the model is `https://huggingface.co/google/mt5-small`

[2]For details about Hugging Face Callbacks, please refer to `https://huggingface.co/docs/transformers/main_classes/callback`

also be used as the baselines without the need for any modification.

6.1.2 Feminine and Masculine Classifiers

As shown in Table 4, the feminine data set contains fewer sentences than the masculine data set. To ensure a balanced class distribution for training the classifier, I used the same amount of sentences for both gender categories, i.e. 45,800 sentences each. For comparison, the classifiers were also trained on the original data sets, i.e. before filtering, and an imbalanced class distribution, i.e. more masculine sentences than feminine ones. Different variations of the classifiers were tracked with Weights & Biases (Biewald, 2020). The results will be shown in Chapter 7.

Regarding the architecture of the classifiers, I adopted the classifier structure of the formality change task from (Yang and Klein, 2021), specifically, a 3-layer causal LSTM (Hochreiter and Schmidhuber, 1997) with a hidden dimension of 512. A dropout layer is appended to the output of each LSTM layer except the last layer, with a dropout probability equal to 0.5. As mentioned by Yang and Klein (2021), the classifiers should use the same vocabulary as the generation models (trimmed mt5-small). I also initialized the embeddings in the classifier from the pre-trained mt5-small, this is not mandatory however, embeddings can be initialized randomly or pretrained with another method.

To train the classifiers, I adapted the scripts from SimpleFUDGE (Kew and Ebling, 2022). Both classifiers were trained for 15 epochs starting with a starting learning rate of 0.001 and a Cosine Annealing (Loshchilov and Hutter, 2017) learning rate scheduler[3] implemented by PyTorch (Paszke et al., 2019).

Feminine and masculine classifiers $\mathcal{B}_f$ and $\mathcal{B}_m$ were trained on the same ParlaMint 2.1 (Erjavec et al., 2021) data sets with the same LSTM architecture. To train $\mathcal{B}_f$, sentences by female speakers are the positive class, whereas in training $\mathcal{B}_m$, sentences by male speakers are the positive class.

6.2 Decoding

Since the gender *female* and *male* are treated as two opposing attributes, I decided to test them separately for this project. The general idea is that for the FUDGE models with the feminine classifier, I only test it on the sentences uttered by female

[3] `https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.CosineAnnealingLR.html`

speakers, and then test the ones with the masculine classifier on the sentences uttered by male speakers.

As introduced in Chapter 4, Yang and Klein (2021) augments the underlying generator model's logits with log probabilities from the classifier. In the experiments, the English–Italian translation models $\mathcal{G}$ and $\mathcal{G}_t$ each were augmented with the feminine and masculine classifiers $\mathcal{B}_f$ and $\mathcal{B}_m$ respectively, yielding four combinations of FUDGE, as illustrated in Figure 1

Considering that Kew and Ebling (2022)'s SimpleFUDGE implements beam search for decoding, while FUDGE only allows greedy search, the SimpleFUDGE scripts were adapted and used for decoding. The beam size was set to 4, the same as the default value in SimpleFUDGE.

6.3 Evaluation

6.3.1 BLEU

As a standard Machine Translation evaluation metric, BLEU (Papineni et al., 2002) was used as the first method to assess FUDGE's performance. To ensure the comparability of the results, I used SacreBLEU (Post, 2018) as the tool to calculate the BLEU scores.

6.3.2 MuST-SHE Gender Translation Evaluation Method

Since BLEU provides a global score about translation "quality" as a whole and does not reflect the systems' ability to produce the correct gender forms (Bentivogli et al., 2020), I also used MuST-SHE Evaluation (Savoldi et al., 2022) to assess the models. MuST-SHE Evaluation assesses the translation of gender terms at two levels of granularity, i.e. word-level POS gender evaluation and chain-level gender agreement evaluation.

Word-level Evaluation

The POS annotation is on the Italian side. With the POS annotation, MuST-SHE performs a fine-grained qualitative analysis of the system's accuracy in producing the target gender-marked words. Savoldi et al. (2022) compute accuracy as the proportion of gender-marked words in the references that are correctly translated by the system. An upper bound of one match for each gender-marked word is applied

sentence_id	correct	wrong	speaker_gender	pos
it-0005	dottor	dottoressa	F	noun
it-0005	un	un'	F	art/prep
it-0005	italoamericano	italoamericana	F	noun

Table 9: An illustration of the POS annotations for example sentence pair (6.1), which includes the correct target word, the one in the opposite gender, the gender of the speaker, and the POS category. In this case, the gender of the speaker is female, but the word *dottor* is referring to a male, hence the correct gender form is masculine instead of feminine.

to prevent rewarding over-generated terms.

To illustrate the POS annotation, for the example sentence (6.1), we will have these three rows from Table 9 recorded in the POS annotation table, including the target word, the word in the opposite gender, the POS category of the word, the gender of the speaker, and the ID of the sentence where it comes from, which, in this case, is it-0005. In this example, the gender of the speaker is female, but the word *dottor* is referring to a male, hence the correct gender form is masculine instead of feminine.

(6.1) *His name was Dr. Pizzutillo, an Italian American, ...*
Il suo nome è Dottor Pizzutillo, un Italoamericano, ...
'Il suo nome è Dottor Pizzutillo, un Italoamericano, ...'

In evaluation, for each annotation, if any one of the correct or the wrong word is found, the respective annotation is in coverage, and then the accuracy is further calculated within the in-coverage annotations. One limitation is that for each word, if the system translates it into a synonym of the target, then it will not be treated as in coverage.

Gender Agreement Evaluation

Agreement annotation is also on the Italian side. Gender Agreement Evaluation inspects agreement chains in the translation. Savoldi et al. (2022) define *coverage* as the proportion of generated chains matching with those annotated in MuST-SHE. Each agreement chain involves several agreement terms, if any of the agreement terms do not appear in the translation, then it will be regarded as *Out of Coverage*, meaning the translation failed to include this agreement chain.

And in the case of all the agreement terms appearing in the translation, Savoldi

et al. (2022) distinguish between these three cases:

1. agreement is respected, and with the correct gender (Correct);
2. agreement is respected, but with the wrong gender (Wrong);
3. both feminine and masculine gender inflections occur together, and thus the agreement is not respected (No).

As an illustration, in an example English–Italian pair (5.1), the agreement chain contains three agreement terms, *Dottor*, *un*, and *Italoamericano*, and each paired with the opposite gender terms: *Dottoressa*, *un'*, and *Italoamericana*.

(6.2) *His name was Dr. Pizzutillo, an Italian American, ...*
Il suo nome è Dottor Pizzutillo, un Italoamericano, ...
'Il suo nome è Dottor Pizzutillo, un Italoamericano, ...'

Table 10 displays a few cases that might be seen during the evaluation. If any of the term pairs are missing in the translation, which means there are only one or two terms that appear (e.g. *Dottor* and *Italoamericano* are in the translation, neither *un* nor *un'* appears), then this agreement chain is treated as out of coverage, such as the first row in Table 10. When all three terms appear:

1. if all of them are correct (*Dottor*, *un*, and *Italoamericano*), then it is an example of a Correct Agreement (the second row in Table 10);
2. if all of them are from the opposite gender terms (*Dottoressa*, *un'*, and *Italoamericana*), it is an example of a Wrong Agreement (the third row in Table 10);
3. if a mix of the correct and opposite gender terms appear, then it is a No Agreement (e.g. the bottom row in Table 10).

In summary, the Correct Agreement and Wrong Agreement both have only one case, while Out of Coverage and No Agreement could have multiple cases.

Nevertheless, MuST-SHE does not evaluate the overall translation fluency. Hence it is not sufficient to be a standalone evaluation method for machine translation. The combination of BLEU and MuST-SHE is necessary.

	Category
Dottor, *Italoamericano*	Out of Coverage
Dottor, *un*, *Italoamericano*	Correct Agreement
Dottoressa, *un'*, *Italoamericana*	Wrong Agreement
Dottoressa, *un'*, *Italoamericano*	No Agreement

Table 10: Different cases of agreement chain evaluation categories. Out of Coverage refers to translation examples that involve less than three terms, e.g. the first row. Correct and Wrong only have one case, either all three terms are correct or all three terms are from the opposite gender, which is the second and the third row. If a mix of gender terms appears, then it is No Agreement, e.g. the bottom row.

7 Results

7.1 Standard and Tagged Baseline

The BLEU scores of the standard baseline and the tagged baseline on both the test sets by female and male speakers are displayed in Table 11. On the test set with sentences by female speakers, the tagged baseline has a BLEU score of 27.5, which is 0.3 more than the standard baseline. On the test set with sentences by male speakers, with a BLEU score of 27.1, the tagged baseline is 0.1 higher than the standard baseline.

7.2 Feminine and Masculine Classifiers

Before and after data filtering

Chapter 5.2 mentioned filtering out sentences that do not contain adjectives or participles. Figure 2 shows the validation accuracy of two feminine classifiers during training. The orange line is the validation accuracy after splitting the utterances into sentences. And then I further filtered out the sentences that do not contain adjectives or participles, whose accuracy is shown as the blue line[1]. They were both trained for 15 epochs. The x-axis is the step size and the y-axis is the validation accuracy. Each dot in the graph represents the validation accuracy of a checkpoint. The processed data has more sentences, which require more steps to traverse the data set, hence the orange line is longer than the blue line. The checkpoints with higher validation accuracy were selected as the respective classifier for decoding. Preprocessing the data (splitting it into sentences), results in an accuracy of 0.59. Filtering out the sentences with no marked gender increases the accuracy of the classifier further to 0.69.

[1]The number of sentences are described in Chapter 5.2 and shown in Table 4

	standard baseline	tagged baseline
feminine	27.2	**27.5**
masculine	27.0	**27.1**

Table 11: The BLEU score of the standard baseline and the tagged baseline on both the test sets by female and male speakers. The tagged baseline has a higher BLEU score in both types of test sets.

	causal LSTM	bidirectional LSTM
BLEU	**26.2**	20.1

Table 12: The BLEU score of tagged FUDGE tested on the sentences by female speakers with the causal LSTM classifier and the bidirectional LSTM classifier. Tagged FUDGE's $\lambda = 5$. The validation accuracy of the two classifiers is shown in Figure 3

Causal LSTM and bidirectional LSTM

Figure 3 shows the changes in the validation accuracy of two feminine classifier architectures during training. The blue line is a causal 3-layer LSTM and the gray line is a 3-layer bidirectional LSTM. The rest of the hyperparameters were the same and they were both trained for 15 epochs. The x-axis is the step size and the y-axis is the validation accuracy. Each dot in the graph represents the validation accuracy of a checkpoint and the checkpoints with higher validation accuracy were selected as the respective classifier for decoding. The checkpoint with 0.69 validation accuracy was selected as the causal LSTM classifier and the checkpoint with 0.81 validation accuracy was selected as the bidirectional LSTM classifier. The performance of these two classifiers will be shown below.

Table 12 shows the BLEU score of tagged FUDGE tested on the sentences by female speakers with the causal LSTM classifier and the bidirectional LSTM classifier. Tagged FUDGE's $\lambda = 5$. The tagged FUDGE with the causal LSTM classifier has a BLEU score of 26.2, 6.1 more than the one with the bidirectional LSTM classifier, which is 20.1.

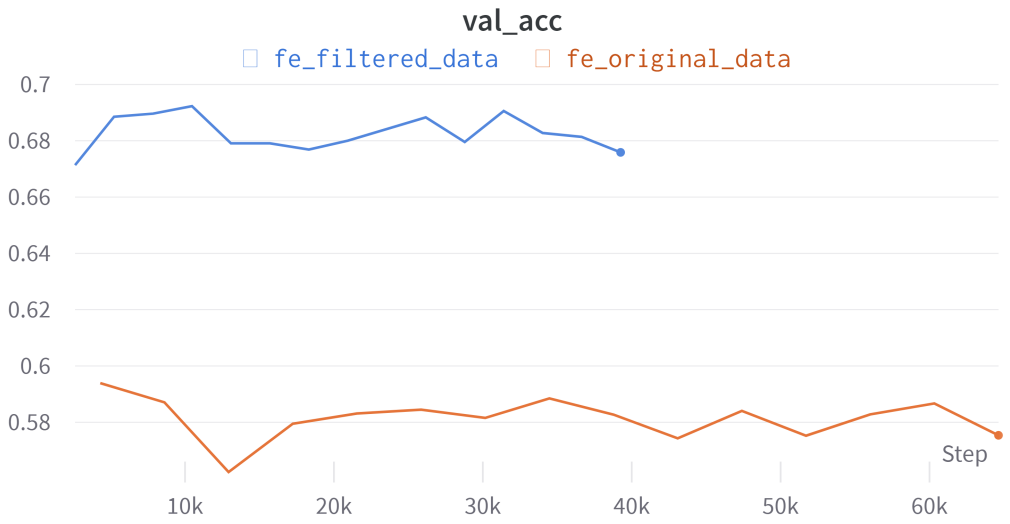


Figure 2: The validation accuracy of two feminine classifiers during training. The blue line was trained on the sentences containing both adjectives and participles. The orange line was trained on the sentences before the filtration. They were both trained for 15 epochs. The x-axis is the step size and the y-axis is the validation accuracy. Each dot in the graph represents the validation accuracy of a checkpoint. The highest validation accuracy of the classifier trained on the filtered data set is 0.69 and the classifier trained on the processed data set is 0.59.

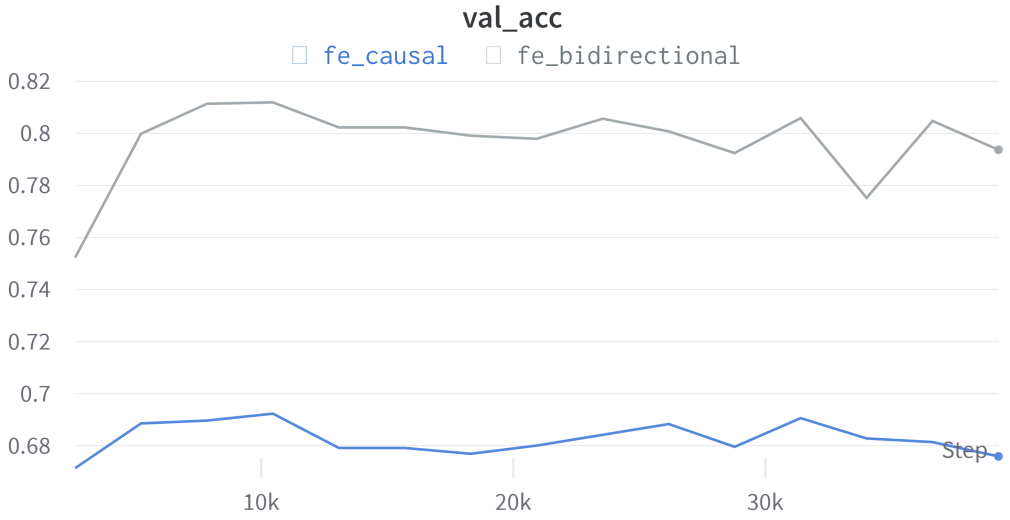


Figure 3: The validation accuracy of two feminine classifier architectures during training. The blue line is a causal 3-layer LSTM and the gray line is a 3-layer bidirectional LSTM. The rest of the hyperparameters were the same and they were both trained for 15 epochs. The x-axis is the step size and the y-axis is the validation accuracy. Each dot in the graph represents the validation accuracy of a checkpoint. The highest validation accuracy of the causal LSTM classifier is 0.69 and the bidirectional LSTM classifier is 0.81.

	standard FUDGE		*tagged FUDGE*	
	feminine	*masculine*	*feminine*	*masculine*
$\lambda = 0$	**27.2**	**27.0**	**27.5**	**27.1**
$\lambda = 1$	27.1	**27.0**	27.3	26.9
$\lambda = 2$	27.0	26.8	27.2	26.9
$\lambda = 3$	26.9	26.7	27.0	26.7
$\lambda = 4$	26.5	26.6	26.6	26.5
$\lambda = 5$	26.2	26.4	26.2	26.5

Table 13: The BLEU score of standard FUDGE and tagged FUDGE tested on both the test sets by female and male speakers, i.e. the four models illustrated in Figure 1. Each model was tested on λ ranging from 1 to 5. When $\lambda = 0$, the classifier does not contribute, hence FUDGE's output is equivalent to the underlying translation models and the baselines.

7.3 BLEU Score of Standard and Tagged FUDGE

Table 13 shows the BLEU score of standard FUDGE and tagged FUDGE tested on both the sentences by female and male speakers, i.e. the four models illustrated in Figure 1. As mentioned in section 4.3, each model was tested on λ ranging from 1 to 5. When $\lambda = 0$, the classifier does not contribute, hence FUDGE's output is the same as the underlying translation models or the baselines.

As displayed in Figure 4, the baselines have the highest BLEU score on both the test sets by female and male speakers. With the increase of λ's value, the BLEU score either does not change or decreases, as depicted in Figure 4.

7.4 MuST-SHE Gender Translation Evaluation

MuST-SHE Gender Translation Evaluation includes **word-level** gender evaluation, which computes the accuracy based on the POS (Parts-Of-Speech) annotations, and **chain-level** gender agreement evaluation, which computes the accuracy of the generated chains.

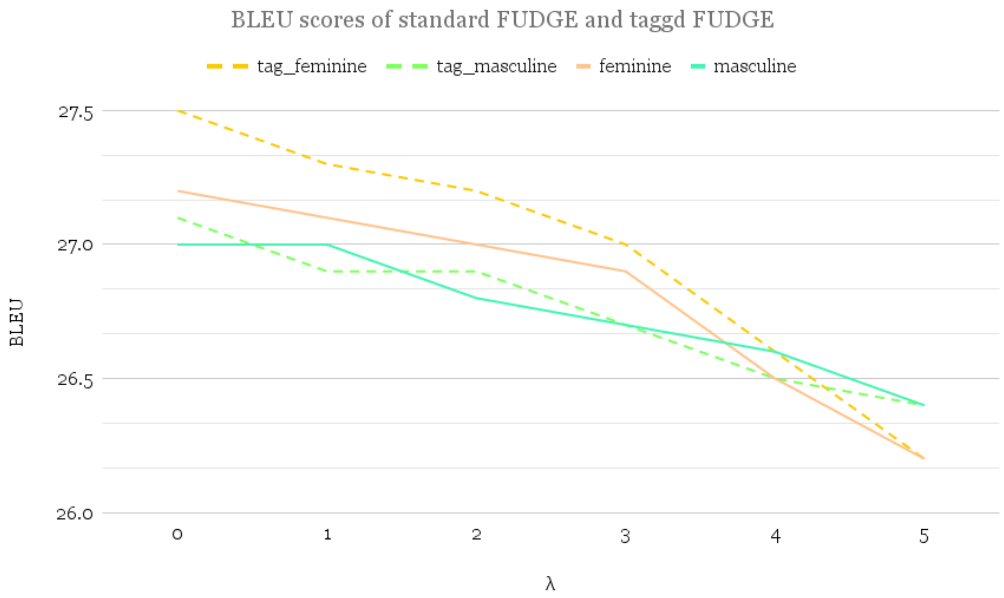


Figure 4: A visualization of Table 13. The variation of the BLEU score with the changes of λ from 0 to 5. When $\lambda = 0$, the BLEU score is the same as the baseline. The x-axis is the values of λ and the y-axis is the BLEU score ranging from 26 to 27.5. *tag_feminine* and *tag_masculine* denote the two models from tagged FUDGE. feminine and masculine denotes the two models from standard FUDGE.

7.4.1 Word-level Gender Evaluation

7.4.1.1 Coverage scores of POS Annotation

Table 14 displays the word-level POS annotation coverage of all six POS categories. As mentioned in section 6.3.2, one POS annotation is in coverage only if either the target word or the word with the opposite gender form appears in the translation. Both the feminine and masculine POS annotation coverage doesn't vary much with the change of λ's value. For feminine POS annotations, standard and tagged FUDGE demonstrate comparable coverage with both a bit more than 38%. And for the masculine POS annotations, they both have around 42% coverage. But the average feminine POS coverage is a bit lower than the masculine POS for both standard FUDGE and tagged FUDGE.

	standard FUDGE		*tagged FUDGE*	
	feminine	*masculine*	*feminine*	*masculine*
baseline	37.6	**42.3**	37.2	**42.4**
$\lambda = 1$	38.2	42.0	37.2	**42.4**
$\lambda = 2$	**38.9**	42.0	**38.2**	42.3
$\lambda = 3$	37.9	**42.3**	38.0	42.3
$\lambda = 4$	37.3	41.9	37.2	42.3
$\lambda = 5$	37.4	41.9	37.9	41.3
average	37.9	42.1	37.6	42.2

Table 14: The word level feminine and masculine **POS coverage** of standard and tagged FUDGE with λ ranging from 1 to 5. The coverage numbers in the first row represent the respective baselines.

7.4.1.2 Translation Accuracy on Open Class POS

Word-level evaluation calculates the accuracy of all six categories of words shown in Table 7. Only the accuracy of the open-class words was selected to display here since open-class words are often marked with the gender of the speaker. First, the results of standard and tagged FUDGE are shown separately in two tables to give an impression of the performance differences between the two series of models. Then we will look at them from a different angle, the translation accuracy of the feminine and masculine forms, to further observe the accuracy differences between the two genders.

Standard FUDGE

Table 15a shows the feminine and masculine open class POS accuracy of standard FUDGE with λ ranging from 1 to 5. The first row displays the accuracy scores tested on the standard baseline.

On the translation of feminine open class POS, when $\lambda = 5$, the standard FUDGE achieves the highest accuracy on all three types, with 71% on verbs, 17.1% on nouns, and 61.4% on descriptive adjectives. And the standard baseline has the lowest accuracy on verbs and descriptive adjectives.

On the translation of masculine open class POS, FUDGE and the baseline have comparable high accuracy scores on nouns and descriptive adjectives. FUDGE's

	feminine			*masculine*		
	Verbs	*Nouns*	*Adj-des*	*Verbs*	*Nouns*	*Adj-des*
baseline	27.4	11.4	35.4	87.8	**97.6**	94.3
$\lambda = 1$	43.7	12.8	42.9	91.4	96.3	**94.4**
$\lambda = 2$	60.6	13.2	61.2	92.9	97.5	94.2
$\lambda = 3$	62.1	10.8	55.1	94.1	97.4	94.1
$\lambda = 4$	70.1	11.8	61.2	**96.9**	97.5	94.1
$\lambda = 5$	**71.0**	**17.1**	**61.4**	96.6	97.5	92.0

(a) **Standard** FUDGE's accuracy on open-class POS

	feminine			*masculine*		
	Verbs	*Nouns*	*Adj-des*	*Verbs*	*Nouns*	*Adj-des*
baseline	27.3	13.5	36.3	94.4	**97.6**	94.1
$\lambda = 1$	39.5	13.2	45.7	94.5	97.5	92.2
$\lambda = 2$	56.3	**20.5**	55.1	95.8	97.5	91.7
$\lambda = 3$	63.6	14.3	61.7	93.1	97.5	92.2
$\lambda = 4$	**67.1**	15.4	64.6	**97.0**	97.3	**96.1**
$\lambda = 5$	62.9	19.0	**66.0**	95.5	97.5	91.8

(b) **Tagged** FUDGE's accuracy on open-class POS

Table 15: The **word-level** feminine and masculine open-class POS translation accuracy with λ ranging from 1 to 5. The first row of each table displays the accuracy scores from the respective baseline. *Adj-des* denotes descriptive adjectives, as discussed in section 5.3.

noun accuracy is 10 percentage points higher than the baseline.

Tagged FUDGE

Having the same layout as Table 15a, Table 15b shows the open class POS accuracy of tagged FUDGE. The first row displays the accuracy scores tested on the tagged baseline.

On the translation of feminine open class POS, FUDGE is more accurate in all three categories than the baseline, 67.1% on verbs, 20.5% on nouns, and 66% on descriptive adjectives. Similar to the standard baseline, the tagged baseline has a lower accuracy on verbs (27.3%) and descriptive adjectives (36.3%).

	standard FUDGE		*tagged FUDGE*	
	feminine	*masculine*	*feminine*	*masculine*
$\lambda = 0$	21.7	23.9	23.0	22.2
$\lambda = 1$	23.7	23.5	23.0	23.1
$\lambda = 2$	**25.0**	23.1	**25.0**	22.2
$\lambda = 3$	23.7	23.1	24.3	23.1
$\lambda = 4$	23.0	**24.4**	24.3	23.1
$\lambda = 5$	21.7	22.2	**25.0**	**24.4**
average	23.1	23.4	24.1	23.0

Table 16: The feminine and masculine **agreement chain coverage** of standard and tagged FUDGE with λ ranging from 1 to 5. When $\lambda = 0$, FUDGE is equivalent to the underlying translation model, hence the coverage numbers in the first row represent baselines.

On the translation of masculine open class POS, again, the baseline and FUDGE both maintain high accuracy on all three open class POS. One thing worth noting is that the tagged baseline's noun accuracy is 6 percentage points higher than the standard baseline.

7.4.2 Chain-level Gender Agreement Evaluation

7.4.2.1 Agreement Chain Coverage

For gender agreement evaluation, the *coverage* is the first metric. As shown in Table 16, the standard and the tagged FUDGE have identical coverage rates on both the feminine and masculine agreement chains, with around 23% to 25%. Tagged FUDGE has a higher average feminine agreement chain coverage than standard FUDGE. Interestingly, the coverage rate of feminine agreement chains is comparable with the masculine.

7.4.2.2 Agreement Chain Accuracy

Standard FUDGE

Among the agreement chains that are in coverage, the accuracy is further evalu-

ated. Table 17a presents the feminine and masculine agreement accuracy of standard FUDGE with λ ranging from 1 to 5. The first row displays the results from the **standard baseline**. As discussed in section 6.3.2, *Correct* means that the agreement is respected with the correct gender, *Wrong* means that the agreement is respected but with the wrong gender, and *No* means that the agreement is not respected.

Standard FUDGE demonstrates a steady increase in feminine agreement chain accuracy with the increase of λ's value and reaches 63.6% accuracy when $\lambda = 5$, almost 20 percentage points more than the baseline. For masculine agreement chains, FUDGE even reduces the *wrong agreement* percentage to 0%.

Tagged FUDGE

With a similar structure, Table 17b shows the accuracy of the tagged FUDGE. The tagged baseline (first row of 17b) performs better than the standard baseline (first row of 17a). Nevertheless, with the increase of λ, standard FUDGE demonstrates more improvement than the tagged FUDGE.

	feminine			*masculine*		
	Correct	*Wrong*	*No*	*Correct*	*Wrong*	*No*
baseline	45.5	36.4	18.2	91.1	3.6	5.4
$\lambda = 1$	52.8	33.3	13.9	94.5	1.8	3.6
$\lambda = 2$	57.9	28.9	**13.2**	94.4	1.9	3.7
$\lambda = 3$	52.8	27.8	19.4	94.4	1.9	3.7
$\lambda = 4$	57.1	20.0	22.9	**96.5**	**0**	**3.5**
$\lambda = 5$	**63.6**	**18.2**	18.2	92.3	**0**	7.7

(a) Gender agreement chain accuracy of **standard** FUDGE

	feminine			*masculine*		
	Correct	*Wrong*	*No*	*Correct*	*Wrong*	*No*
baseline	48.6	37.1	**14.3**	**96.2**	**0**	3.8
$\lambda = 1$	45.7	34.3	20	94.4	1.9	3.7
$\lambda = 2$	52.6	31.6	15.8	94.2	1.9	3.8
$\lambda = 3$	**56.7**	**27.0**	16.2	94.4	1.9	3.7
$\lambda = 4$	51.3	32.4	16.2	**96.2**	**0**	3.7
$\lambda = 5$	44.7	34.2	21.1	94.7	1.8	**3.5**

(b) Gender agreement chain accuracy of **tagged** FUDGE

Table 17: The gender **agreement** evaluation results of standard and tagged FUDGE on both the feminine and masculine agreement chains with λ ranging from 1 to 5. The first row of each table displays the results from the respective baseline. As discussed in section 6.3.2, *Correct* means that the agreement is respected with the correct gender, *Wrong* means that the agreement is respected but with the wrong gender, and *No* means that the agreement is not respected. The numbers represent the percentage of each case and it was calculated separately for the feminine and masculine agreement chains.

8 Discussion

8.1 Standard Baseline and Tagged Baseline

The data presented in Table 11 shows that the tagged baseline improves more on the utterances by female speakers. The advantage of adding a gender tag to the English source side is more noticeable for feminine sentences, which is also where most of the gender bias occurs.

Table 18 displays the BLEU score on the English–Italian direction of the baseline (equivalent to the standard baseline) and the gender-enhanced NMT systems (equivalent to the tagged baseline) from Vanmassenhove et al. (2018). The gender-enhanced NMT system has a higher BLEU of 0.29 than the baseline, which is similar to the tagged baseline's +0.3 BLEU improvement on the feminine-only test set.

Note that the results shown here from Vanmassenhove et al. (2018) are on a data set that contains both feminine and masculine sentences. They tested the systems on feminine-only and masculine-only data sets and claimed that the biggest BLEU score improvement is observed on the feminine test set.

A greater improvement in the utterances by female speakers is expected. As discussed in Chapter 5, the female gender is under-represented in the training data, hence the baseline model tends to translate the sentence into the masculine form.

8.2 Feminine and Masculine Classifiers

Before and after data filtration

As illustrated in Figure 2, the classifier trained on the filtered data set has higher accuracy, despite the fact that it was only trained on 90,000 sentences compared to the one trained on processed data which contains 567,000 sentences. It validates the assumption that the adjectives and participles provide a stronger gender signal during the training of the classifiers, while the data set before filtration contains too

	baseline	tagged NMT
EN-IT	31.46	**31.75**

Table 18: The BLEU score of the baseline (equivalent to standard baseline) and gender-enhanced NMT systems (equivalent to the tagged baseline) on the English–Italian direction from Vanmassenhove et al. (2018).

many gender-neutral sentences that might confuse the classifier. The two classifiers have the same LSTM architecture, so the deciding factor is the difference in the training data. Hence the classifier trained on filtered data is chosen to proceed with the experiments.

Causal LSTM and bidirectional LSTM

As illustrated in Figure 3, the bidirectional LSTM classifier has a higher validation accuracy than the causal LSTM classifier. But during decoding, the tagged FUDGE with the bidirectional LSTM classifier has a much lower BLEU score than the one with the causal LSTM classifier. The higher validation accuracy indicates that the bidirectional LSTM is more suitable to classify the gender of the sentences because of having access to both the start and the end of the sentence. However, the decoding process generates one token at a time, which does not allow the classifier to attend both directions at the same time. But the generation flow matches the input flow of the causal classifier. So the performance of the causal LSTM classifier is better, which is also the reason the causal LSTM was chosen to be the classifier architecture.

8.3 BLEU Score of FUDGE

As shown in Table 13, the BLEU scores of both standard and tagged FUDGE decreases with the increase of λ, though the changes are not dramatic. The main reason might be that the classifiers' training set is too small due to the filtering of sentences that do not contain adjectives or participles, so FUDGE's good performance in correcting the gender-marked term is at the cost of translation fluency. Next, let us review some of FUDGE's translation examples, which might shed some light on this issue.

English source	[FEMALE] The internet is a **medium** ...
Italian reference	Internet è un **mezzo** ...
FUDGE	
$\lambda = 3, 4, 5$	Internet è un **media** ...
$\lambda = 1, 2$	Internet è un **medio** ...
Baseline	
Both	Internet è un **medio** ...

Table 19: An overcorrected translation example of tagged FUDGE on a sentence by a female speaker with λ ranging from 1 to 5. When λ equals 1 and 2, FUDGE translates the English word *medium* into a masculine word *medio*, which refers to the middle finger as a noun. When λ increases, FUDGE uses a morphologically similar feminine noun, *media*, meaning "average value". But it neglects the article, *un*, remaining in the masculine form.

Overcorrection

One possible cause of the decrease is that when λ's value gets bigger, the classifier has a stronger influence on the word choice and unnecessarily changes some words that are not relevant to the gender of the speaker. And these words are usually in the opposite gender form of the speaker's gender. As illustrated in the translation example of the word *medium* in Table 19, the gender of the speaker is female, but the correct translation should be a masculine noun, *mezzo*. When λ equals 1 and 2, FUDGE generates a masculine word *medio*, which refers to "the middle finger´´ as a noun and is not a common word in Italian. When λ increases, FUDGE uses a morphologically similar feminine noun, *media*, meaning "average value". But it neglects the article, *un*, remaining in the masculine form. As shown in this example, FUDGE tends to select words in feminine form even when it is not necessary.

Inadequate Reference

Another reason for the BLEU score decreasing could be that there is only one Italian reference sentence for each English source sentence in the Europarl-Speaker-Information (Vanmassenhove and Hardmeier, 2018) data set. But given a sentence, there could be different correct translations with different word choices. During translation, FUDGE usually follows the same sentence structure as the underlying translation model, except that it modifies the translation of some words and phrases.

In this translation example, the tag and the gender of the speaker do not match, i.e.

English source	[FEMALE] I am **delighted** with the work of the Ombudsman, ...
Italian reference	Sono **soddisfatto** del lavoro del Mediatore, ...
FUDGE	
$\lambda = 1 - 5$	Sono **lieta** del lavoro del Mediatore, ...
Baseline	
Both	Sono **lieto** del lavoro del Mediatore, ...

Table 20: A correct translation example of tagged FUDGE on a sentence by a female speaker with λ ranging from 1 to 5. The gender tag on the English source side is "[FEMALE]", while the word *soddisfatto* in the reference is masculine. Tagged FUDGE generates the correct feminine form *lieta*. And the two baselines translate it into the masculine form *lieto*.

the tag is female, but the gender-marked word in the sentence is in masculine form, but FUDGE manages to translate it into the correct feminine form. As displayed in Table 20, the English word *delighted* is translated into *soddisfatto* but the tag is *FEMALE*. But FUDGE translates it into *lieta*, the correct feminine form. And both the standard and tagged baseline translate it into *lieto*. This example also demonstrates that FUDGE usually has the same word choice as the baseline, but changes its gender form.

Inaccurate Translation

On the flip side, FUDGE may overlook some phrases in translation. In Table 21, while FUDGE preserves the translation of the phrase *you will agree* when $\lambda = 1, 2, 3$, it omits *che lei concorderà* when λ's value is higher, equals 4 or 5.

Looking at the gender-marked word *sure*, the two baselines use the masculine form *sicuro*. FUDGE also uses *sicuro* when $\lambda = 1$, which indicates the feminine classifier does not make a difference in this case. When $\lambda = 2, 3$, FUDGE still uses the same word but with the feminine form *sicura*, and when $\lambda = 4, 5$, FUDGE changes the word choice of the model and selects the same word as the Italian reference with the feminine form *certa*.

English source	[FEMALE] I am **sure** you will agree, Madam President, ...
Italian reference	Sono **certa** che sarà d'accordo, signora Presidente, ...
FUDGE	
$\lambda = 4, 5$	Sono **certa**, signora Presidente, ...
$\lambda = 2, 3$	Sono **sicura** che lei concorderà, signora Presidente, ...
$\lambda = 1$	Sono **sicuro** che lei concorderà, signora Presidente, ...
Baseline	
Both	Sono **sicuro** che lei concorderà, signora Presidente, ...

Table 21: A translation example of tagged FUDGE on a sentence by a female speaker with λ ranging from 1 to 5. When $\lambda = 4$ and 5, tagged FUDGE generates the same word *certa* as the Italian reference sentence. When $\lambda = 2$ and 3, tagged FUDGE generates a synonym of *certa*, *sicura*, with the correct feminine form. And it generates the word in masculine form when $\lambda = 1$, the same as the two baselines.

8.3.1 Additional Translation Example

Here is another translation example from the Europarl-Speaker-Information test set. First, if we only look at the word *delighted* and its translations in Table 22, the feminine tagged FUDGE translates *delighted* into the correct feminine form *lieta*, and both the standard and tagged baselines translate it into the masculine form *lieto*. Considering the sentence translation as a whole, there are four versions:

1. the Italian reference from Europarl-Speaker-Information;
2. the tagged FUDGE translation when $\lambda = 1, 2, 4, 5$;
3. the tagged FUDGE translation when $\lambda = 3$ and the tagged baseline;
4. the standard baseline

Among these different versions, the reference is more accurate on translating *are heading* into *proceda*, while the others are missing a verb. For the phrase *in this respect*, the translation when $\lambda = 1, 2, 4, 5$ is more accurate than other versions, even the Italian reference, because it correctly translates it into *in questo ambito* while the others simply ignore it. Here the reader may notice that the translation of the tagged baseline and tagged FUDGE are the same when $\lambda = 3$, while they are different when $\lambda = 1, 2$. But in general, we would expect that the smaller λ's value, the more similar the translations of FUDGE and the underlying translation model,

English source	[FEMALE] I am **delighted** that we are all heading in the same direction in this respect.
Italian reference	Sono davvero **lieta** che si proceda nella stessa direzione.
FUDGE	
$\lambda = 1, 2, 4, 5$	Sono **lieta** che siamo tutti nella stessa direzione in questo ambito.
$\lambda = 3$	Sono **lieta** che siamo tutti nella stessa direzione.
Baseline	
Tagged	Sono **lieto** che siamo tutti nella stessa direzione.
Standard	Sono **lieto** che siamo tutti in questa direzione nella stessa direzione.

Table 22: A translation example of tagged FUDGE on a sentence by a female speaker with λ ranging from 1 to 5. All of the FUDGE models generate the word in the correct feminine form *lieta*, the same as the reference, while the two baselines all generate the masculine form *lieto*.

as shown in the above translation example.

8.4 MuST-SHE Gender Translation Evaluation

8.4.1 Word-level Gender Evaluation

8.4.1.1 Coverage

From Table 14, we can see that FUDGE increased the feminine POS coverage by around 1 percentage point and almost have no influence on masculine POS coverage. The possible reason is that FUDGE didn't change the word choice in translation, but uses a different gender form.

8.4.1.2 Accuracy

Translation Accuracy of the Baselines

Table 23 displays the first lines from Table 15a and 15b, which are from the standard and the tagged baseline. For the feminine open class POS, the tagged baseline performs better on nouns and descriptive adjectives and has almost the same per-

	feminine			*masculine*		
	Verbs	*Nouns*	*Adj-des*	*Verbs*	*Nouns*	*Adj-des*
standard	**27.4**	11.4	35.4	87.8	**97.6**	**94.3**
tagged	27.3	**13.5**	**36.3**	**94.4**	**97.6**	94.1

Table 23: The feminine and masculine open class POS accuracy of the standard and the tagged baseline. *Adj-des* denotes descriptive adjectives, as discussed in section 5.3.

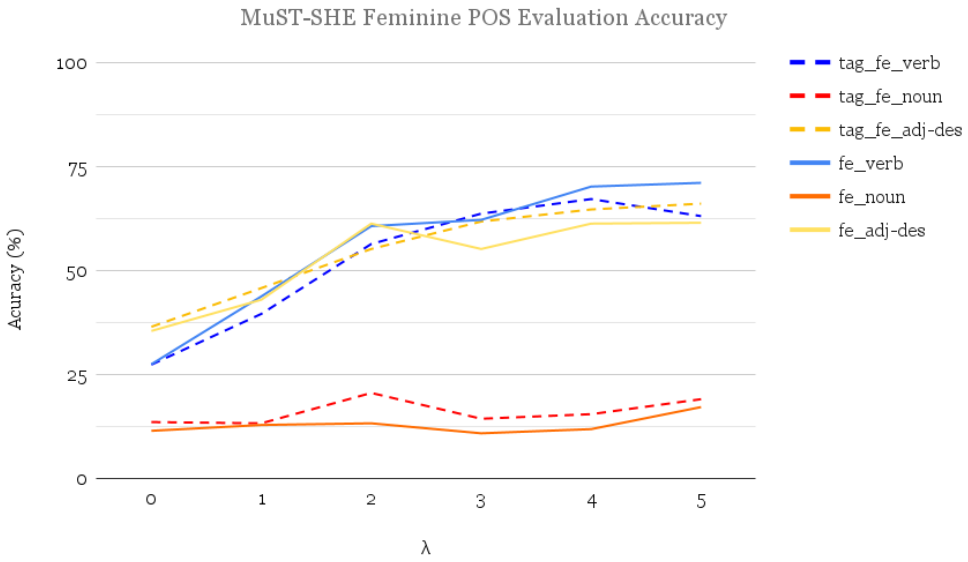
formance as the standard baseline. For the masculine open class POS, two baselines have similar accuracy on nouns and descriptive adjectives, while the tagged baseline performs better on verbs. These results validate the findings by Vanmassenhove et al. (2018), that the tagged translation system performs better on feminine sentences than the masculine sentences.

Figure 5a depicts the variation of accuracy on the feminine open class POS in Table 15a and 15b and Figure 5b on the masculine open class POS. In both charts, the x-axis is the value of λ and the y-axis is the accuracy. The results from tagged FUDGE are in dashed lines and standard FUDGE solid lines. Verbs are in blue, nouns are in red, and descriptive adjectives (adj-des) are in yellow. To reflect the differences in feminine and masculine open class POS translation accuracy, the y-axes in both figures are the same, from 0% to 100% accuracy.

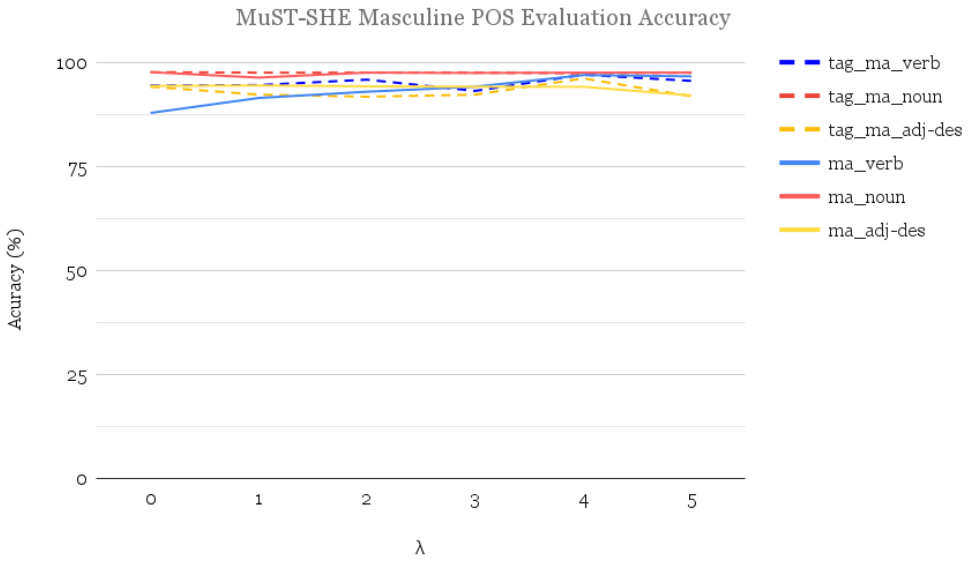
Translation Accuracy of the Feminine Form

As shown in Figure 5a, standard FUDGE and tagged FUDGE share an identical pattern on the accuracy of feminine open class POS. With the increase of λ, the accuracy of verbs increases 40 percentage points from approximately 30% to 70%, descriptive adjectives increase 30 percentage points from around 35% to 65%.

However, the accuracy of nouns improves by only around 5 percentage points, which is much lower than verbs and descriptive adjectives. The reason is that when verbs and descriptive adjectives are in a gendered form, they usually agree with the gender of the speaker, while a masculine noun has an equal chance to be in a sentence uttered by a female speaker or a male speaker, the same for a feminine noun. Hence it's difficult for the classifier to predict which one to use in which case. This is also the reason for FUDGE's overcorrection as seen in Table 19.



(a) Feminine open class POS translation accuracy



(b) Masculine open class POS translation accuracy

Figure 5: Visualization of Table 15. Figure 5a and 5b show the feminine and masculine POS accuracy respectively. The x-axis is the value of λ and the y-axis is the accuracy. The results from tagged FUDGE are in dashed lines and standard FUDGE solid lines. Verbs are in blue, nouns are in red, and descriptive adjectives (adj-des) are in yellow. When $\lambda = 0$, FUDGE is equivalent to the underlying translation model, hence the data points represent the accuracy of the baselines.

Translation Accuracy of the Masculine Form

From Figure 5b, we clearly notice that all three open-class categories have high ac-

	standard FUDGE		*tagged FUDGE*	
	feminine	*masculine*	*feminine*	*masculine*
word-level POS	**37.9**	**42.1**	**37.6**	**42.2**
agreement chain	23.1	23.4	24.1	23.0

Table 24: The average coverage rates of both the feminine and masculine **word level POS** and **agreement chain** for standard and tagged FUDGE with λ ranging from 0 to 5.

curacy on the masculine open-class POS and do not vary much with the changes of λ, except for a slight increase in verbs. Besides, the differences between standard and tagged FUDGE on masculine open-class POS translation are almost indistinguishable. Together with the translation examples above, the results show that the NMT models usually treat the masculine form as the default translation option.

Summary

To summarize the Word-level Evaluation, both standard and tagged FUDGE demonstrate huge improvement in the accuracy of feminine open class POS, though the improvement of standard FUDGE is a bit higher. But even after applying FUDGE, there's still a gap in translation accuracy between the feminine and masculine forms.

8.4.2 Chain-level Gender Agreement Evaluation

8.4.2.1 Agreement Chain Coverage

If we compare the average coverage of word-level POS annotation and the agreement chains, as presented in Table 24, we may see that the coverage rate in the word level is much higher than in the agreement chain evaluation.

One reason is that the requirement is more strict when considering whether an agreement chain is in coverage. As discussed in section 6.3.2, it applies the POS annotation coverage requirement to each agreement term in the chain, only if all terms meet the requirement, the chain will be in coverage.

Another reason is that for each agreement term, only one correct translation and one opposite-gender translation are provided. When the model translates the English term into a synonym of the provided agreement terms, even if the translation is

adequate, it is still not in coverage.

8.4.2.2 Agreement Chain Accuracy

For the agreement chains that are in coverage, the next step is to evaluate their accuracy. Figure 6a visualizes the feminine gender agreement accuracy from Table 17a and 17b. The results of the baseline are represented by the data points when $\lambda = 0$. The standard FUDGE is in solid lines and the tagged FUDGE is in dashed lines. The correct agreement is in blue, the wrong is in red and the no agreement is in yellow. And the masculine gender agreement accuracy is illustrated in Figure 6b.

Feminine Gender Agreement Accuracy

Figure 6a demonstrates that tagged FUDGE's *correct agreement* percentage is going up and down, and the *wrong agreement* percentage drops and then increases. While standard FUDGE has a more steady performance improvement with an increasing correct percentage and a decreasing wrong percentage. The *no agreement* percentage doesn't vary much with λ's value for both models.

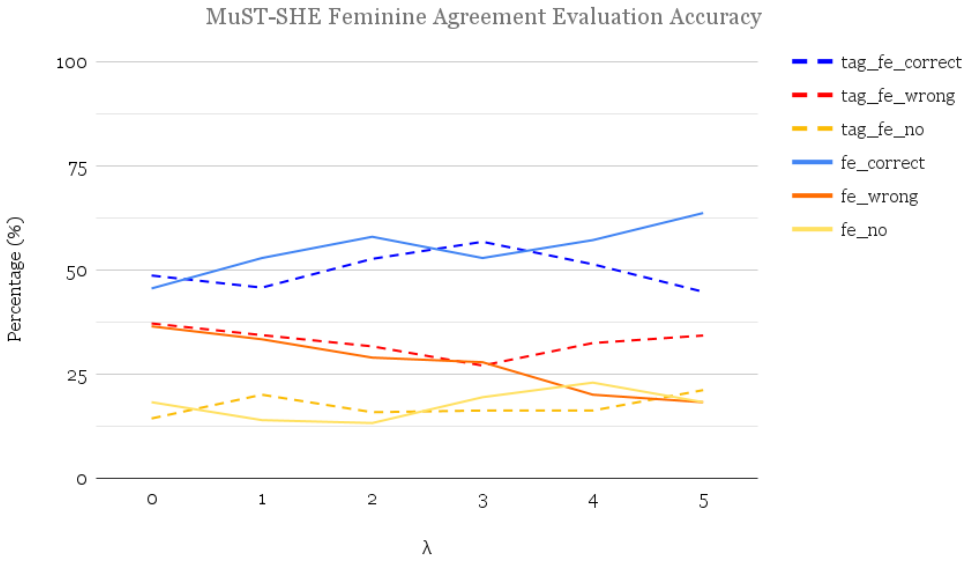
The improvement of feminine agreement accuracy is less significant. The reason is that agreement evaluation is closely related to word-level evaluation. In the feminine open-class POS translation, FUDGE demonstrates a low accuracy on nouns, even though it is much more accurate on verbs and descriptive adjectives. An agreement chain has a high chance of containing nouns as its agreement terms, and the rule requires the model to correctly translate every term. But FUDGE does not improve noun translation accuracy, hence it undermines the agreement accuracy.

Masculine Gender Agreement Accuracy

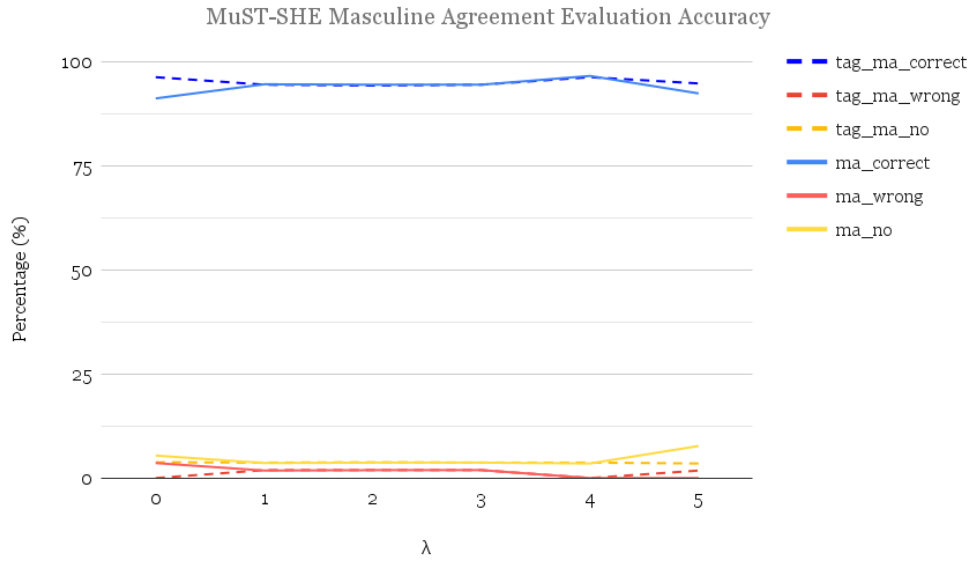
For the results of masculine agreement accuracy are displayed in Figure 6b, standard and tagged FUDGE both maintain a high correct percentage and pretty low wrong and no percentages. Compared to the standard baseline, standard FUDGE increases five percentage points in *correct agreement*, from 91.1% to 96.5%, and the wrong percentage drops from 3.6% to 0%. A good performance on masculine agreement evaluation is expected, considering FUDGE's high accuracy on all three open-class POS categories.

Summary

Overall, for chain-level gender agreement evaluation, standard FUDGE consistently improves performance on both feminine and masculine agreements, while tagged FUDGE exhibits an unstable performance on the feminine agreement chains.



(a) Feminine agreement chain translation accuracy



(b) Masculine agreement chain translation accuracy

Figure 6: Visualization of Table 17. Figure 6a and 6b show the feminine and masculine agreement translation accuracy respectively. The x-axis is the value of λ and the y-axis is the percentage. The results from tagged FUDGE are in dashed lines and standard FUDGE solid lines. Correct agreements are in blue, wrong agreements are in red, and no agreements are in yellow. When $\lambda = 0$, FUDGE is equivalent to the underlying translation model, hence the data points represent the accuracy of the baselines.

9 Conclusion

Contribution

The main contribution of this work is that it utilizes a controlled text generation method, Future Discriminators for Generation (FUDGE), to mitigate gender bias in NMT. The experiments were conducted in English and Italian and the results exhibited significant improvement in feminine gender terms accuracy with concrete translation examples.

One thing that distinguishes this work from others is that it pinpointed the type of gender bias to tackle and designed the experiments accordingly. First, considering *Speaking As* dimension of the gender bias mostly occurs in first-person sentences, the training sets of the underlying translation model and the classifiers are both parliament proceedings, which contain a high proportion of first-person utterances. During the evaluation, apart from BLEU, we also incorporate a novel fine-grained gender translation evaluation metric, the MuST-SHE gender translation evaluation method, which evaluates gender translation on a word level and agreement level. The combination of these two evaluation metrics enables a qualitative and quantitative assessment of the model.

The results show that FUDGE has a slightly lower BLEU score than the baseline, but demonstrates promising outcomes on MuST-SHE evaluation. On the word level, FUDGE significantly increases the feminine gender term accuracy across multiple POS categories. On the gender agreement chain level, FUDGE also improves the correct agreement percentage.

Future Research

One possible direction for future research would be to investigate whether FUDGE can also mitigate other types of gender biases, e.g. stereotypical gender bias. In this case, WinoMT would be suitable as an evaluation benchmark.

In this work, I used separate models with the respective classifiers for test sets by female and male speakers, another potential idea for future research might be to unify these models, which would require integrating the feminine and masculine

classifiers into one system and finding suitable λs to balance them.

References

M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn, S. Ortiz Rojas, L. Pla Sempere, G. Ramírez-Sánchez, E. Sarrías, M. Strelec, B. Thompson, W. Waites, D. Wiggins, and J. Zaragoza. ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.417. URL `https://aclanthology.org/2020.acl-main.417`.

L. Bentivogli, B. Savoldi, M. Negri, M. A. Di Gangi, R. Cattoni, and M. Turchi. Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.619. URL `https://aclanthology.org/2020.acl-main.619`.

L. Biewald. Experiment Tracking with Weights and Biases, 2020. URL `https://www.wandb.com/`. Software available from wandb.com.

J. Coates. *Women Men and Language : A Sociolinguistic Account of Gender Differences in Language.* Pearson Longman, Harlow, England, 2004.

G. G. Corbett. *Agreement.* Cambridge University Press, 2006.

S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *International Conference on Learning Representations.* OpenReview.net, 2020. URL `https://openreview.net/forum?id=H1edEyBKDS`.

M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota, June 2019. Association for

Computational Linguistics. doi: 10.18653/v1/N19-1202. URL https://aclanthology.org/N19-1202.

E. Dinan, A. Fan, L. Wu, J. Weston, D. Kiela, and A. Williams. Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.23. URL https://aclanthology.org/2020.emnlp-main.23.

P. Eckert and J. R. Rickford. *Style and Sociolinguistic Variation*. Cambridge University Press, 2001.

A. El-Kishky, V. Chaudhary, F. Guzmán, and P. Koehn. CCAligned: A Massive Collection of Cross-lingual Web-Document Pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.480. URL https://www.aclweb.org/anthology/2020.emnlp-main.480.

T. Erjavec, M. Ogrodniczuk, P. Osenova, N. Ljubešić, K. Simov, V. Grigorova, M. Rudolf, A. Pančur, M. Kopp, S. Barkarson, S. Steingrímsson, H. van der Pol, G. Depoorter, J. de Does, B. Jongejan, D. Haltrup Hansen, C. Navarretta, M. Calzada Pérez, L. D. de Macedo, R. van Heusden, M. Marx, Ç. Çöltekin, M. Coole, T. Agnoloni, F. Frontini, S. Montemagni, V. Quochi, G. Venturi, M. Ruisi, C. Marchetti, R. Battistoni, M. Sebők, O. Ring, R. Darģis, A. Utka, M. Petkevičius, M. Briedienė, T. Krilavičius, V. Morkevičius, R. Bartolini, A. Cimino, S. Diwersy, G. Luxardo, and P. Rayson. Linguistically Annotated Multilingual Comparable Corpora of Parliamentary Debates ParlaMint 2.1, 2021. ISSN 2820-4042. URL http://hdl.handle.net/11356/1431. Slovenian language resource repository CLARIN.SI.

C. Escolano, M. R. Costa-jussà, J. A. R. Fonollosa, and M. Artetxe. Multilingual Machine Translation: Closing the Gap between Shared and Language-specific Encoder-Decoders. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 944–948, Online, Apr. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.80. URL https://aclanthology.org/2021.eacl-main.80.

J. Ficler and Y. Goldberg. Controlling Linguistic Style Aspects in Neural Language Generation. In *Proceedings of the Workshop on Stylistic Variation*,

pages 94–104, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4912. URL https://aclanthology.org/W17-4912.

B. Friedman and H. Nissenbaum. Bias in Computer Systems. *ACM Transactions on Information Systems*, 14(3):330–347, jul 1996. ISSN 1046-8188. doi: 10.1145/230538.230561. URL https://doi.org/10.1145/230538.230561.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

C. Hardmeier, M. R. Costa-jussà, K. Webster, W. Radford, and S. L. Blodgett. How to Write a Bias Statement: Recommendations for Submissions to the Workshop on Gender Bias in NLP. *Computing Research Repository*, arXiv:2104.03026, 2021. URL https://arxiv.org/pdf/2104.03026.pdf.

S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020. doi: 10.5281/zenodo.1212303.

P. Jwalapuram, S. Joty, and Y. Shen. Pronoun-Targeted Fine-tuning for NMT with Hybrid Losses. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2267–2279, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.177. URL https://aclanthology.org/2020.emnlp-main.177.

N. S. Keskar, B. McCann, L. Varshney, C. Xiong, and R. Socher. CTRL - A Conditional Transformer Language Model for Controllable Generation. *Computing Research Repository*, arXiv:1909.05858, 2019. URL https://arxiv.org/pdf/1909.05858.pdf.

T. Kew and S. Ebling. Target-Level Sentence Simplification as Controlled Paraphrasing. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 28–42, Abu Dhabi, United

Arab Emirates (Virtual), Dec. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.tsar-1.4.

P. Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, Sept. 13-15 2005. URL https://aclanthology.org/2005.mtsummit-papers.11.

B. Krause, A. D. Gotmare, B. McCann, N. S. Keskar, S. Joty, R. Socher, and N. F. Rajani. GeDi: Generative Discriminator Guided Sequence Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.424. URL https://aclanthology.org/2021.findings-emnlp.424.

H. J. Levesque, E. Davis, and L. Morgenstern. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press, 2012. ISBN 9781577355601.

S. Loáiciga, S. Stymne, P. Nakov, C. Hardmeier, J. Tiedemann, M. Cettolo, and Y. Versley. Findings of the 2017 DiscoMT Shared Task on Cross-lingual Pronoun Prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 1–16, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4801. URL https://aclanthology.org/W17-4801.

I. Loshchilov and F. Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

OpenAI. GPT-4 Technical Report. *Computing Research Repository*, arXiv:2303.08774, 2023. URL https://arxiv.org/pdf/2303.08774.pdf.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito,

M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

M. Post. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

N. Roberts, D. Liang, G. Neubig, and Z. Lipton. Decoding and Diversity in Machine Translation. In *NeurIPS 2020 Workshop on Resistance AI*, 2020. URL https://www.amazon.science/publications/decoding-and-diversity-in-machine-translation.

R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2002. URL https://aclanthology.org/N18-2002.

B. Savoldi, M. Gaido, L. Bentivogli, M. Negri, and M. Turchi. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874, 2021. doi: 10.1162/tacl_a_00401. URL https://aclanthology.org/2021.tacl-1.51.

B. Savoldi, M. Gaido, L. Bentivogli, M. Negri, and M. Turchi. Under the Morphosyntactic Lens: A Multifaceted Evaluation of Gender Bias in Speech Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1807–1824, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.127. URL https://aclanthology.org/2022.acl-long.127.

P. Schachter and T. Shopen. *Language Typology and Syntactic Description. Vol. 1: Clause Structure*. Cambridge University Press, 2 edition, 2007. doi: 10.1017/CBO9780511619434.

R. Sennrich. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL `https://aclanthology.org/E17-2060`.

R. Sennrich, B. Haddow, and A. Birch. 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, Aug. 2016a. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL `https://aclanthology.org/P16-1162`.

R. Sennrich, B. Haddow, and A. Birch. Controlling Politeness in Neural Machine Translation via Side Constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California, June 2016b. Association for Computational Linguistics. doi: 10.18653/v1/N16-1005. URL `https://aclanthology.org/N16-1005`.

D. Stahlberg, F. Braun, L. Irmen, and S. Sczesny. Representation of the Sexes in Language. *Social Communication*, pages 163–187, 01 2007.

G. Stanovsky, N. A. Smith, and L. Zettlemoyer. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1164. URL `https://aclanthology.org/P19-1164`.

J. Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

E. Vanmassenhove and C. Hardmeier. Europarl Datasets with Demographic Speaker Information. In *EAMT 2018 - Proceedings of the 21st Annual*

Conference of the European Association for Machine Translation. European Association for Machine Translation, 2018.

E. Vanmassenhove, C. Hardmeier, and A. Way. 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, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1334. URL https://aclanthology.org/D18-1334.

E. Vanmassenhove, D. Shterionov, and A. Way. Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland, Aug. 2019. European Association for Machine Translation. URL https://aclanthology.org/W19-6622.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

K. Yang and D. Klein. FUDGE: Controlled Text Generation With Future Discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.276. URL https://aclanthology.org/2021.naacl-main.276.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. 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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003.