Universität Zürich[UZH]

Gender-Aware Neural Machine Translation

Master's Thesis in Informatics
submitted by
Benjamin Suter
Student ID: 09-729-005

Department of Informatics
University of Zurich
Supervisor: Prof. Dr. Martin Volk
Submission Date: 03/02/2022

Abstract

This thesis addresses the issue of gender bias in machine translation. It presents a simple yet effective approach to controlling gender morphology in the target language. It focuses on gender morphology in the 1st and 2nd person (speaker and addressee) and suggests using gender tags at the sentence level to direct the model to the desired gender.

Its main contributions are the creation of two gender-annotated parallel corpora for English–Russian and English–French, and several experiments analyzing the effect of gender tagging on translation quality.

Experimental results show that the use of appropriate gender tags leads to a significant improvement in translation quality (at least +2.14 BLEU), with a particularly high improvement for sentences referring to a female person (up to +9.97 BLEU).

Zusammenfassung

*Diese Arbeit befasst sich mit dem Thema des Gender Bias in der maschinellen Übersetzung und stellt eine einfache, aber effektive Methode zur Kontrolle der Genusmorphologie in der Zielsprache vor. Im Zentrum der Arbeit steht die Genusmorphologie in der 1. und 2. Person (Sprecher*in und Adressat*in) und es wird vorgeschlagen, Genus-Tags auf Satzebene zu verwenden, um das Modell auf das gewünschte Genus zu lenken.*

Der Hauptbeitrag dieser Arbeit besteht in der Erstellung von zwei genusannotierten parallelen Korpora für Englisch–Russisch und Englisch– Französisch sowie in der Durchführung mehrerer Experimente, in denen die Auswirkung des Genus-Taggings auf die Übersetzungsqualität analysiert wird.

Die Versuchsergebnisse zeigen, dass die Verwendung geeigneter Genus-Tags zu einer signifikanten Verbesserung der Übersetzungsqualität (mindestens +2.14 BLEU) führt. Eine besonders deutliche Verbesserung wird bei Sätzen erzielt, die sich auf eine weibliche Person beziehen (bis zu +9.97 BLEU).

Contents

1 Introduction

1.1 Problem description

In many languages, when referring to a person, it is necessary to modify certain words depending on whether the person is male or female. In French, for example, *I'm happy* is expressed either as *Je suis heureux* if the speaker is male, or as *Je suis heureuse* if the speaker is female.

When translating a sentence from a language with gender to a language without gender, the gender information contained in the original sentence is simply omitted in the translation. The following example shows two French sentences differing only in gender and their identical translation into English:

m: *Je suis heureux.*
f: *Je suis heureuse.* → *I'm happy.*

In contrast, when translating in the opposite direction, there are at least two different, equally correct translations for a single source sentence:

I'm happy. → m: *Je suis heureux.*
f: *Je suis heureuse.*

This one-to-many mapping presents a difficulty for machine translation systems. Although the English and French sentences above are translations of each other, their information content is not the same. The French translation must render information about the gender of the person that is absent from the original English sentence and therefore cannot be inferred from it. Without additional information, there is no reason to favor one gender over the other in translation, but that is exactly what standard machine translation models are forced to do. As a result, the gender in the

output is random, unpredictable and additionally may vary from sentence to sentence, potentially causing a lot of tedious post-editing work.

Importantly, the choice of a gender is not simply a stylistic choice, such as, for example, that between synonyms, but it affects the meaning of the resulting sentence in a significant way. Both of the above variants are correct translations of the English sentence, but only one of them is correct in a given context. A text with gender forms that do not match the situation (e.g., in movie subtitles) will be perceived as wrong and confusing.

The lack of gender information in the input cannot be compensated for by better learning algorithms or a larger collection of training data. The solution must instead be to give the user the option of passing this gender information along with the input text. From an implementation standpoint, the easiest way to pass this information along with the input text is to append or prepend it literally to the input text in the form of tags indicating the desired gender. For example, if we assume that the tags `<m>` and `<f>` denote male and female, respectively, then the input and output of a machine translation system might look as follows:

`<m>`	*I'm happy.*	$\rightarrow$	*Je suis heureux.*
`<f>`	*I'm happy.*	$\rightarrow$	*Je suis heureuse.*

Such a model would be able to handle the gender decision systematically, translating each sentence in accordance with the given gender tag in the input. However, the above example makes an undue simplification: it assumes that gender is an attribute of a whole sentence, while it is, in fact, an attribute of a person. Several persons may be mentioned in a single sentence, not necessarily all of the same gender. Therefore, if several persons are mentioned in a sentence, it is necessary to specify the gender of each person separately.

One possible approach could be to apply the tags to individual words rather than entire sentences, for example: *The doctor <f> greeted the patient <m>*. However, this solution is relatively complex, as users need to know which words in the target language need to be gendered, and potentially tedious, as the original input text needs to be edited to insert the tags at the right place. Indeed, manually post-editing a translated sentence containing inappropriate gender forms may often be quicker than pre-editing the original sentence.

Interestingly, the vast majority of previous work on gender in machine translation (see Chapters 2 and 3) has focused exclusively on third-person gender, that is, gender in sentences like *My friend is a lawyer* or *The doctor greeted the patient*. Gender of the 1st person (speaker) and the 2nd person (addressee) has received far less attention, although in many languages, such as French or Russian, gender agreement is equally prevalent in all persons.

At the same time, the 1st and 2nd person differ from the 3rd person in one essential respect: they are always unique in a sentence, i.e. there is always at most one 1st person (*'I'*) and one 2nd person (*'you'*).[1] For this reason, it is not necessary to place a gender tag next to a specific word (as in 3rd person). Instead, it is sufficient to define the gender of, e.g., the 1st person globally for the whole sentence. In other words, for the 1st and 2nd person, it is possible to specify their gender using sentence-level tags, as initially desired. All that is needed is a tag set that refers to the speaker and one that refers to the addressee.

For example, if we assume that the tags <1m> and <1f> denote a male or a female speaker, respectively, and <2m> and <2f> a male or a female addressee, then the input and output of a machine translation system may look as follows:

<1m>	<2m>	*I'm sure you're tired.*	→	*Je suis sûr que tu es fatigué.*
<1f>	<2f>	*I'm sure you're tired.*	→	*Je suis sûre que tu es fatiguée.*
<1m>	<2f>	*I'm sure you're tired.*	→	*Je suis sûr que tu es fatiguée.*
<1f>	<2m>	*I'm sure you're tired.*	→	*Je suis sûre que tu es fatigué.*

Gender is not the only phenomenon where certain words in a sentence depend on the persons referred to. Another common phenomenon is that of honorifics. In languages with honorifics, the choice of certain words depends on the social hierarchy or distance between the speaker and the addressee. In many Indo-European languages, for example, a distinction is made between formal and informal language, and the two levels of language are distinguished by different pronouns and verb forms in reference to the addressee, e.g., *tu* (*you*, informal) and *vous* (*you*, formal) in French. This distinction is often referred to as T–V distinction, named after the Latin 2nd person pronouns *tu* (singular) and *vos* (plural).

[1] In the plural of the 1st and 2nd person (*'we'* or plural *'you'*), it is a group of people instead, but the group itself is still unique.

French has a T–V distinction, and hence, for example, *That's your book* is expressed either as *C'est ton livre* in informal language, or as *C'est votre livre* in formal language. Again, when translating a sentence from a language without honorifics such as English to a language that requires the choice of a politeness level, there are at least two different, equally correct translations:

That's your book.	→	informal:	*C'est ton livre.*
		formal:	*C'est votre livre.*

As with gender, the politeness level in French cannot be derived from the original English sentence and therefore must be passed to the machine translation system in addition to the input text. Since politeness expression apply to full sentences, it is possible to do this again with two mutually exclusive sentence-level tags that denote informal and formal language, respectively. For example, if we assume that the tag <T> denotes informal language and the tag <V> denotes formal language, then the input and output of a machine translation model might look as follows:

<T> *That's your book.*	→	*C'est ton livre.*
<V> *That's your book.*	→	*C'est votre livre.*

In fact, this solution was proposed and implemented earlier by Sennrich et al. [2016a], and their research results for English-to-German show that appending such tags to the input text does indeed allow to control the level of politeness in the translated sentence, and that this also improves the overall translation quality in terms of the standard evaluation metric BLEU [Papineni et al., 2002].

Despite some similarities, though, gender and politeness also have important differences: While expressions of politeness refer to the social hierarchy or distance between speaker and addressee in a given situation, and each individual may be addressed both formally and informally depending on the situation, gender is an integral part of a person's identity and does not usually change from situation to situation. Consequently, if a machine translation model has a bias towards a certain politeness level (e.g., formal), this may result in inappropriate translations of some texts; however, it will not affect any particular group of people. On the other hand, if a machine translation model has a bias towards one of the genders, it discriminates people of the neglected gender.

Moreover, the choice between informal and formal language is often associated with different text domains. For example, a speech in a parliament is usually made in formal language, while a letter to a good friend is written in informal language. Texts in formal and informal language may therefore have differences in words or phrases that go beyond what is grammatically necessary, and this will likely be reflected in a machine translation model that takes politeness levels into account. Gender, on the other hand, cannot or should not be associated with a particular text domain and, consequently, with a particular language use, and therefore translation differences that go beyond what is grammatically necessary are undesirable.

In languages that have both grammatical gender and a T–V distinction, the 2nd person is affected by both phenomena. This is the case with many European languages, e.g. French, Russian or German. Because of this overlap, it is important for a model to consider both phenomena together to be useful in practice. The combination of gender tags and politeness tags makes it possible to configure up to eight different translations into French from a single English input sentence, all of them equally correct in isolation:

`<1m>`	`<2m>`	`<T>`	*I'm glad you're my friend.*	→	*Je suis heureux que tu sois mon ami.*
`<1f>`	`<2f>`	`<T>`	*I'm glad you're my friend.*	→	*Je suis heureuse que tu sois mon amie.*
`<1m>`	`<2f>`	`<T>`	*I'm glad you're my friend.*	→	*Je suis heureux que tu sois mon amie.*
`<1f>`	`<2m>`	`<T>`	*I'm glad you're my friend.*	→	*Je suis heureuse que tu sois mon ami.*
`<1m>`	`<2m>`	`<V>`	*I'm glad you're my friend.*	→	*Je suis heureux que vous soyez mon ami.*
`<1f>`	`<2f>`	`<V>`	*I'm glad you're my friend.*	→	*Je suis heureuse que vous soyez mon amie.*
`<1m>`	`<2f>`	`<V>`	*I'm glad you're my friend.*	→	*Je suis heureux que vous soyez mon amie.*
`<1f>`	`<2m>`	`<V>`	*I'm glad you're my friend.*	→	*Je suis heureuse que vous soyez mon ami.*

The goal of the present thesis is to develop a gender-aware machine translation model for the language pairs English–Russian and English–French in accordance with the previous discussion. The model should be able to correctly interpret tags added to the input string that indicate the gender of the speaker, the gender of the addressee, or the desired politeness level, and produce correct translations given these side constraints. It should have the following properties:

- The tags are on the sentence level, i.e., it is not necessary for the user to identify and annotate specific words.

- It is possible to specify more than one, and up to three constraints at the same time, for instance the gender of the speaker and the gender of the addressee.
- It is possible to use the model without tags.
- Tags can be passed to the model regardless of the content of the sentence; it is the task of the model to decide whether or not the tags have an impact on the translation.

The models are evaluated in several experiments. The key questions are: 1) are the models able to generate translations that take into account one or more side constraints, and 2) how does the use of tags affect the overall translation quality in terms of BLEU?

1.2 Structure of the thesis

This thesis is organized as follows: Chapter 2 explains the motivation for this work and shows that current commercial systems such as DeepL and Google Translate exhibit a strong gender bias.

Chapter 3 discusses previous work related to the topic of this thesis. This includes research assessing gender bias in machine translation, research on the general topic of additional constraints on the output in machine translation, as well as research on the specific topic of gender-aware machine translation.

Chapter 4 introduces the methods used to create the gender-annotated parallel corpora for English–Russian and English–French. Furthermore, it describes the model architecture and its hyperparameters as well as the training procedure.

Chapter 5 describes several experiments used to evaluate the effects of gender tagging on translation quality, and Chapter 6 presents the results of these experiments.

Chapter 7 provides a more detailed discussion of the experimental results and of potential applications of gender-tagging in machine translation, as well as a discussion of remaining issues of the proposed method and potential future work. The thesis closes with a short conclusion in Chapter 8.

2 Motivation

2.1 Gender bias in machine translation

When using automatic translation services, I have repeatedly found that while the overall translation quality is very good, I often have to correct inappropriate gender or politeness forms in the output. As outlined in Chapter 1, when models translate from a non-gendered language to a gendered language, they are forced to choose one of the genders for references to persons in the text, and the same is true for the politeness level. There is only a fifty percent chance a priori that the random choice of the model matches the correct gender or politeness level in a given context, which ultimately means that in about half of all sentences with gender and politeness marking, correction of the inappropriate gender and politeness forms is necessary as a post-editing step.

Input	DeepL	m/f	Google	m/f	Yandex	m/f
I was	Я был	m	я был	m	Я был	m
I said	Я сказал	m	я сказал	m	Я сказал	m
I wanted	Я хотел	m	я хотел	m	Я хотел	m
I thought	Я думал	m	я думал	m	Я думал	m
I was able	Я смог	m	Я мог	m	Я был в состоянии	m
I made	Я сделал	m	я сделал	m	Я сделал	m
I knew	Я знал	m	я знал	m	Я знал	m
I saw	Я видел	m	Я видел	m	Я видел	m
I talked	Я говорил	m	Я говорил	m	Я разговаривал	m
I found	Я нашел	m	я нашел	m	Я нашел	m

Table 2.1: Translations of ten frequent past-tense verbs from English to Russian with DeepL, Google Translate and Yandex Translate and the resulting gender in Russian (11/12/2021).

Input	DeepL	m/f	Google	m/f	Yandex	m/f
I went	Je suis allé	m	je suis allé	m	Je suis allé	m
I arrived	Je suis arrivé	m	Je suis arrivé	m	Je suis arrivé	m
I became	Je suis devenu	m	je suis devenu	m	Je suis devenu	m
I was born	Je suis né	m	je suis né	m	Je suis né	m
I stayed	Je suis resté	m	je suis resté	m	Je suis resté	m
I returned	Je suis revenu	m	je suis rentré	m	Je suis revenu	m
I went out	Je suis sorti	m	Je suis sorti	m	Je suis sorti	m
I came	Je suis venu	m	je suis venu	m	Je suis venu	m
I entered	Je suis entré	m	Je suis entré	m	Je suis entré	m
I fell	Je suis tombé	m	je suis tombé	m	Je suis tombé	m

Table 2.2: Translations of ten past-tense verbs from English to French with DeepL, Google Translate and Yandex Translate and the resulting gender in French (11/12/2021).

To get an idea of which gender the models tend to choose, I conducted a black box test with three popular machine translation services: DeepL[1], Google Translate[2], and Yandex Translate[3]. To this end, I extracted the ten most frequent past-tense verbs from the English–Russian training corpus (see Chapter 4) and translated these verbs from English into Russian in the 1st person singular. For each translation, I noted the gender in the resulting translation. Table 2.1 shows the results of this experiment. The results are surprisingly consistent: all three services produced the masculine form for all ten verbs. As a double-check, I conducted the same experiment with translation from English to French. In French, only verbs that take the auxiliary verb *être* (*to be*) exhibit gender morphology. Therefore, I selected ten verbs that take *être* and translated their English equivalent into French. The results are reported in Table 2.2. Again, all three services produced the masculine form for all ten verbs.

This anecdotal evidence of a strong preference for male gender forms in commercial machine translation systems is confirmed by several research papers, as will be discussed in more detail in Chapter 3. The gender bias in these services has implications for practice because it has the effect that texts referring to females require more post-editing than texts referring to males, and it is, for example, more expensive to produce an adequate translation for texts by a female author than for those by a male author.

[1] https://www.deepl.com
[2] https://translate.google.com
[3] https://translate.yandex.ru

But more than that, commercial systems often also ignore gender information that *is* available in the source sentence. For example, DeepL translates the French sentence *Je suis arrivée à Paris* (*I [female] arrived in Paris*) into the Italian sentence *Sono arrivato a Parigi* (*I [male] arrived in Paris*). This is obviously a case of gender bias, but it is also simply a wrong translation.

These observations impressively illustrate the need for better control of gender morphology in machine translation, and this work aims to take a step in that direction.

2.2 Gender-neutral language

In recent years, there has been increasing interest in establishing gender-neutral forms of language use in English and other gender languages. A well-known example is the widely accepted use of singular *they*[4] in place of *he* or *she* in English. Considering these developments, it is worth asking whether gender-neutral forms should be included in this work, in addition to masculine and feminine forms.

Many of the world's languages are inherently gender-neutral, with no grammatical gender in nouns, adjectives, verbs, or even pronouns. A survey of 257 different languages published in the World Atlas of Language Structures[5], a large linguistic database, found that more than half of the languages studied were gender-neutral. This large group of languages includes, for example, Finnish, Hungarian, Turkish and Indonesian.

Other languages exhibit gender morphology to varying degrees. Gender morphology is especially common among Indo-European and Afro-Asiatic languages. English also belongs to the group of gender languages, as it requires gender distinction in 3rd person pronouns (*he* and *she*), something that is not the case in truly gender-neutral languages.

In languages such as Romance and Slavic, however, a much larger proportion of the vocabulary than in English is affected by gender morphology. It is therefore much more difficult to establish grammatical patterns to refer to a person in a gender-neutral way. This is especially true in spoken language, where typographical solutions are not an option. While there are

[4] https://en.wikipedia.org/wiki/Singular_they
[5] https://wals.info/chapter/30

some emerging proposals (e.g., the gender-neutral pronoun *iel* in French),[6] none of them has yet found wide application.

For this reason, gender-neutral forms will not be considered further in this thesis. It may be noted, however, that the approach proposed in this work can be easily extended to gender-neutral forms by adding a corresponding tag, assuming sufficient training data is available.

[6] https://dictionnaire.lerobert.com/definition/iel

3 Related Work

3.1 Assessment of gender bias

Gender in machine translation has received much attention in recent years. One line of research focuses on the detection, quantification and assessment of gender bias. An up-to-date literature review on assessment of gender bias in machine translation can be found in Savoldi et al. [2021].

Several papers analyze gender bias in commercial machine translation services. Prates et al. [2019] focus on translation of 3rd person singular pronouns (*he*/*she*) from gender-neutral languages (such as Turkish, Hungarian, and Finnish) into English. They prepare sentences in the form *S/he is X*, where *X* is either an adjective or a profession (e.g., *S/he is a teacher*), and translate these sentences into English using Google Translate. Their results show that Google Translate exhibits a strong tendency to choose the masculine pronoun.

Cho et al. [2019] perform a similar analysis, focusing on translations from Korean into English. In addition to Google Translate, they also examine Naver Papago[1] and Kakao Translator[2]. Their results show that all three translation services have a preference for masculine pronouns.

Similarly, Rescigno et al. [2020] assess the translation of occupation terms and adjectives from English into Italian, French and Spanish in sentences with non-3rd person subjects (e.g., *I am a pianist*). Their evaluation of translations from Google Translate, Microsoft Bing Translator[3], and DeepL shows that all three services give preference to the male gender to varying degrees.

[1] https://papago.naver.com
[2] https://translate.kakao.com
[3] https://www.bing.com/translator

Another line of research is the development of challenge test sets and evaluation protocols for assessing gender bias. Stanovsky et al. [2019] create a test set of 3888 English sentences. Each sentence contains two occupational terms and a 3rd person pronoun, and the semantics of the sentence allow to infer the gender of one of the people involved based on the pronoun. For example, in the sentence *The doctor asked the nurse to help her in the procedure*, it can be inferred that the doctor is female and therefore needs to be translated into the feminine form if there exists a gender distinction in the target language. This test set, referred to as WinoMT, is inspired by Zhao et al. [2018] which introduced a similar monolingual challenge for coreference resolution. WinoMT does not contain reference translations in other languages. Instead, sentences are translated using the model under test, then words are aligned between the source sentences and their translation, and finally heuristic rules are applied to determine whether the gender of the word in question is the expected one. Evaluation of various machine translation systems shows that all of them perform poorly on this task, especially in cases where the gender assignment is anti-stereotypical.

The WinoMT challenge test set has been employed in several papers (Costa-jussà et al. [2020], Kocmi et al. [2020], Bergmanis et al. [2020], Basta et al. [2020], Vamvas and Sennrich [2021]). Costa-jussà et al. [2021] provide recordings of the sentences in WinoMT to enable assessment of gender bias in speech translation (WinoST). Similar to WinoMT, Escudé Font and Costa-jussà [2019] create an English-Spanish test set of 1000 parallel sentences in a similar format, where each sentence is constructed based on a predefined template and contains an occupational term and a 3rd person pronoun.

Other benchmarks for assessing gender differences rely on real-world data rather than controlled sentence templates. Habash et al. [2019] select 12 000 English-Arabic sentence pairs from the OpenSubtitles 2018 corpus [Lison et al., 2018] that contain a 1st person pronoun. They manually annotate each of these sentence pairs with the gender expressed in the Arabic translation (male, female, or none). In addition, for all translations labeled as male or female, they manually create a translation into the other gender.

Bentivogli et al. [2020] focus on speech translation and release the MuST-SHE benchmark. It consists of 1062 samples (audio, transcript, and translation) for English-Italian and 1074 samples for English-French. The

test set is based on MuST-C [Di Gangi et al., 2019], a multilingual corpus with data from TED talks. It consists of segments that require the translation of at least one English gender-neutral word into the corresponding masculine or feminine target word. For each segment, they manually create a second reference translation in the other gender.

Finally, Google recently published a dataset[4] for studying gender bias in machine translation that consists of translated biographies from Wikipedia. Its focus lies on 3rd person pronouns and on long-distance coreference resolution over multiple sentences.

Vamvas and Sennrich [2021] propose a general, reference-free approach for evaluation of lexical disambiguation errors, which they call contrastive conditioning. They first translate both the original sentence and two variants with added disambiguating words using the model under test. For example, for the sentence *The assistant asked the doctor if she needs any help* (in which the doctor is female), they create two contrasting variants by replacing the word *doctor* once by *female doctor* (correct) and once by *male doctor* (incorrect). Then they use an evaluation model that evaluates whether the translation of the original sentence is closer to the translation of the correctly disambiguated variant than to the incorrectly disambiguated variant.

Other research has looked at whether the model architecture can lead to bias. Vanmassenhove et al. [2019] identify the loss of lexical richness and diversity as a general problem in machine translation. They run several experiments with different architectures (statistical models, recurrent models, transformers) for English–French and English–Spanish and find that the models indeed increase the probability of frequent words and decrease the probability of less frequent words, i.e., they amplify the bias present in the data. This also applies to the translation of gender-neutral words in English (e.g., doctor) into gendered variants in French and Spanish, creating a gender bias that is stronger than that in the original data.

Roberts et al. [2020] analyze beam search and find that it underpredicts female gender pronoun as compared to sampling based on the distribution in the data. They also find that focusing on improving BLEU scores results in models having lower translation diversity than humans.

Costa-jussà et al. [2020] assess the accuracy of gender forms in multilingual models. They experiment with both a shared encoder and decoder

[4] https://ai.googleblog.com/2021/06/a-dataset-for-studying-gender-bias-in.html

for all languages and separate encoders and decoders per language, and come to the conclusion that separate coders and decoders are beneficial for gender accuracy.

In summary, research on assessing gender bias has focused on two different settings in which it can occur. One of them is the translation of sentences in which the correct gender of a given word can be inferred from the context (e.g., in WinoMT). In this setting, there is a correct and an incorrect solution, and the gender bias can be measured as the proportion of incorrect gender choices. The other setting is the translation of sentences where the context does not provide information about the correct gender (e.g. *I am a pianist*). Here, both genders are correct, but a gender bias may be observed if one of the genders is selected much more frequently than the other.

To my knowledge, no bias-free result has been reported for either setting and regardless of the test set used and the model tested. This confirms the observations in Section 2.1 and demonstrates that gender bias is an unresolved issue.

3.2 Side constraints in machine translation

Side constraints in machine translation have been introduced in Sennrich et al. [2016a]. They propose to append tags to the input sequence to control linguistic attributes such as the politeness level, tense, or the gender and number of discourse participants. In their work, they test their approach for politeness levels in automatic translation from English to German. They apply two sentence-level tags, one indicating a formal level and one an informal level, and report an increase of +1.4 BLEU points with oracle tagging on a random test set and an increase of +3.2 BLEU points on sentences containing a 2nd person pronoun in the English source sentence.

The tagging approach has since become widely accepted for imposing constraints on translation output. Yamagishi et al. [2016] employ two tags to control the voice (active, passive) in translation from Japanese into English and report an increase of +0.73 BLEU points. Similarly, Feely et al. [2019] employ three tags (informal, polite, formal) to control for honorifics when translating from English into Japanese, and see an increase between +0.3 and +1.5 BLEU points, depending on the dataset.

Takeno et al. [2017] employ tags containing an integer to control the number of words in the output sequence and observe an increase of +0.9 BLEU points in oracle experiments. Similarly, in a more recent work, Lakew et al. [2019] use three tags (short, normal, long) to control the length of the output sequence and report successful application with no degradation in translation quality.

Johnson et al. [2017] use tags in Google's multilingual machine translation model to specify the desired target language. They further report that this approach enables zero-shot translation, i.e., translation between language pairs that were never seen during training. Caswell et al. [2019] apply tags in the area of back-translation. They use a single tag to inform the model that these samples are from a different source. They report an improvement of +1.71 BLEU compared to the baseline model without back-translated samples and an improvement of +0.13 BLEU compared to noised, but untagged back-translation samples. Stergiadis et al. [2021] combine this approach with a second set of tags to control the text domain in translation from English to French. They report improvements between +0.64 and +2.10 BLEU for this multidimensional tagging.

While the tagging approach is most widely used for imposing side constraints, other approaches have also been proposed. Kobus et al. [2017] compare the tagging approach for domain adaptation with a word embedding modification approach. In this latter approach, each word embedding is concatenated with a second embedding that encodes the text domain (e.g., medical). Since the text domain is a sentence-level feature, each token in a sequence has the same domain embedding vector concatenated to its word embedding. Conceptually, this may also be viewed as a form of tagging, where a 'tag' in vector form is appended to the word embeddings, rather than a literal tag to the input string. They find that the embedding approach outperforms tagging in oracle experiments with a BLEU difference of up to +0.92. However, it remains an open question whether this approach also works for multiple constraints at the same time.

In their work on multilingual models, Fan et al. [2021] add a special token to the decoder indicating the target language instead of adding a tag to the input sequence. The advantage of this approach is that the encoder does not see the target language and therefore is bound to encode the input sentence in a target-language independent form.

Finally, Schioppa et al. [2021] address the important topic of multiple side constraints at the same time and propose additive vector-valued inter-

ventions. For each attribute to control, they define an intervention vector which is added to all outputs of the encoder. Simultaneous control for multiple attributes is achieved via a linear combination of each attribute vector. If no control over a particular attribute is desired, its intervention vector is simply set to the null vector. They apply their approach to the translation from English to German and from English to Japanese and simultaneously control the sequence length, the politeness level, and monotonicity (i.e., the proximity of word order between the source and target sequences). In oracle experiments, they observe an improvement over the tagging approach of +0.44 BLEU for German and of +0.08 BLEU for Japanese.

In contrast to tagging, this approach allows to control the output with continuous-valued side constraints. As far as gender is concerned, this is not necessary because morphological gender is discrete. Nevertheless, another advantage of this approach is that it offers the possibility to define a ranking of the importance of different side constraints through different weighting of the intervention vectors, something that is not easily possible with the tagging approach. However, the improvement achieved over the tagging approach is relatively small. In conclusion, there is as yet no method that is clearly superior to tagging, and tagging remains a competitive method for imposing side constraints.

3.3 Gender-specific machine translation

In agreement with the observation in Section 3.1, approaches to mitigating gender bias in machine translation fall broadly into two groups: those that attempt to reduce gender bias within the model (e.g., to improve performance on challenges such as WinoMT), and those that aim to give the user control over the resulting gender. The second group of works is more relevant to the current work, but I briefly discuss the other approaches as well.

Zmigrod et al. [2019] focus on languages with rich morphology (Spanish and Hebrew). They propose a method for converting between masculine-inflected and feminine-inflected noun phrases. In short, the method consists of analyzing a sentence with a syntax parser (including morphological analysis), identifying relevant noun phrases, and reinflecting each gendered word in the noun phrase into the other gender using a word-level re-

inflection model. This method can be used to synthetically generate counterfactual samples that can be used for data augmentation during training.

Saunders and Byrne [2020] explore another method of creating counterfactual samples for morphology-rich languages. They use a small list of gendered words in English (such as *he* or *she*) and create a variant in which these words are gender-swapped. Then they translate these sentences into the target language to create additional parallel samples that are used to fine-tune the model. Their results are inconclusive.

Choubey et al. [2021] focus on improving gender translation accuracy on unambiguously gendered inputs (e.g., WinoMT) and propose gender-filtered self-training. In this method, an initial model is first trained on gender-biased data. Then, samples from a monolingual source language corpus are translated using this model, and if the gender in the translation is correct, the sample and its translation are used as an additional synthetic parallel sample. Finally, the model is retrained from scratch using both the original parallel samples and the synthetic samples. They report an improvement in the gender accuracy on the WinoMT and MuST-SHE test sets using this method.

Rabinovich et al. [2017] aim at preserving author traits in translation. They propose using separate machine translation models for each gender to preserve gender-specific language. To this end, they apply various methods to identify the gender of the author for the samples in the training data, including analysis of given names. They observe a slight decrease in the BLEU score with this approach.

Addressing the problem of anaphora resolution, Voita et al. [2018] propose a modification to the transformer architecture [Vaswani et al., 2017] which allows to pass the previous sentence as context. The source sentence and the context sentence are first encoded in a separate encoder module. Then, an attention layer is used in combination with a gating function to produce a contextual representation of the source sentence. They report an improvement of +0.6 BLEU compared to simply concatenating the two sentences. Nevertheless, Basta et al. [2020] still apply the concatenation method for the translation from English to Spanish and achieve an improvement of +1.09 BLEU over the baseline model without context.

Turning to the second group of works that aim to give the user control over the resulting gender, there are two main approaches: post-processing and input tagging. A special case, however, is Moryossef et al. [2019]. Taking advantage of the fact that machine translation models are able to

perform anaphora resolution, they propose a black-box context injection method to control gender in the output without changing the underlying model. To this end, they add to each source sentence a phrase such as *she said to them* or *he said to her*, indicating both the speaker and the addressee of the sentence. Translating these modified sentences with Google Translate and then removing the redundant translated prefix from the output results in a performance gain of up to +2.3 BLEU. However, a disadvantage of this simple approach is that it is not always possible to identify and remove the redundant prefix.

Habash et al. [2019] follow the post-processing approach and focus on 1st person gender in translations from English to Arabic. They experiment with both rule-based and neural gender reinflection and report improvements over the unmodified raw translations, but also the introduction of new errors.

Similarly, Google has announced in a blog post[5] that it employs a post-processing approach to create translations for both genders from a gender-neutral input sentence. To date, however, the service is very limited and works only for selected languages and only if the input consists of no more than a single sentence.

Turning to tagging approaches, Kuczmarski and Johnson [2018] describe the use of two sentence-level tags (masculine, feminine) to control the gender of the 3rd person pronoun in English when translating from a language with a gender-neutral 3rd person pronoun such as Turkish. The paper is of theoretical nature and does not present experimental results.

Vanmassenhove et al. [2018] exploit the fact that the Europarl corpus [Koehn, 2005] contains metadata about the speaker and create a large parallel corpus with annotated speaker information for 20 language pairs. This metadata is used to train a model with two input tags (male, female). To my knowledge, this is the first application of gender tagging together with Elaraby et al. [2018] (see below). Experimental results for translation from English into 10 different languages show an increase of up to +1.44 BLEU in the case of French, but also slight decrease for some languages (e.g., -0.19 BLEU for Spanish).

In a similar direction, Gaido et al. [2020] manually annotate the MuST-C speech translation corpus [Di Gangi et al., 2019] with the gender of the speaker. They test the tagging approach for speech translation for

[5] https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html

English-to-French and English-to-Italian. Their results show that tagging the speaker's gender has no significant effect on BLEU values compared to the baseline, yet significantly improves the accuracy of gender-specific inflections. They also analyze what happens when the vocal characteristics of the input and the gender tag do not match, and observe a significant drop in performance in this scenario, suggesting that the model relies on both.

Bergmanis et al. [2020] analyze the effect of annotating the target gender on the word level (masculine, feminine, none). They experiment with translating from English into five Indo-European languages. They first apply a morphological tagger to the target language and then a word alignment tool to the parallel sentences, and finally tag the source tokens with the target gender. They report an improvement of up to +4.5 BLEU with oracle tagging.

Similarly, Saunders et al. [2020] propose gender tags on the word level. However, unlike in the above work, only the word of interest is explicitly tagged (e.g., *the developer <F>*). They experiment with WinoMT and tag the occupational terms with their reference gender. Although the gender tag is in theory redundant in the WinoMT setting (because it can be inferred from the context), they observe higher gender accuracy (and no significant effect on BLEU) when using the tag.

In the papers discussed so far, gender tagging is limited to a single attribute, either 1st person gender or 3rd person gender. Elaraby et al. [2018] go one step further and use two tag sets that allow to simultaneously control both the gender of the speaker and the gender of the addressee. They take the English-Arabic sub-corpus of OpenSubtitles [Lison and Tiedemann, 2016] and annotate the corpus with the gender of the speaker and the gender of the addressee, using rules based on both part-of-speech tags and surface forms. They first train the model on the untagged data and then fine-tune it on the tagged data. They report an improvement over the baseline model of +0.58 BLEU on the total test set and +2.14 BLEU on the gender-tagged subset in oracle experiments.

During training they only use tags for sentence pairs in which a relevant gender form is present on the target side. This experimental setup leads to a mismatch between training conditions and real-world conditions, since the model expects tags only when they are relevant for translation. Therefore, during application they first run a classifier that analyzes the input sentence and decides whether to keep the gender tags or remove them if

they are not needed for the translation. This solution is not very elegant, because the classifier may introduce errors. In addition, the approach lacks a solution for the reverse case, i.e., for the case when no gender tags are passed but the Arabic translation requires gendering of certain words.

Their work is closest to this one. However, despite some similarities, the present work goes further in many respects. First, it includes a third tag set to control for politeness. There is no morphological politeness distinction in Arabic, which makes this unnecessary in Arabic. Anyway, the addition of the politeness tag is interesting because while the gender of the speaker and the gender of the addressee are independent of each other, the politeness level and the gender of the addressee both refer to the same person, and thus the same tokens are often affected by both tags.

Second, while they only report BLEU scores and do not analyze whether there is actually an improvement in correct gender inflections, I consistently analyze both BLEU scores and gender accuracy. Third, in addition to oracle experiments, I conduct several other experiments, including random tagging experiments and experiments with tag prediction in triangular translation (see Chapter 5 for more details).

4 Method

4.1 Data

4.1.1 Raw datasets

I use data from the OpenSubtitles2018 corpus [Lison et al., 2018][1] as a training corpus. OpenSubtitles is a parallel corpus of movie subtitles available for a wide range of language pairs. Movie subtitles usually contain a lot of direct speech and thus a relatively high number of 1st and 2nd person references, which makes them an interesting data set for this thesis. In addition, movie subtitles are also a potential application area for gender-aware machine translation. Machine translation combined with post-editing is already widely used in professional subtitle translation, and Etchegoyhen et al. [2014] report that having to repeatedly correct politeness levels in the translation can be frustrating for post-editors, and and it can be assumed that the same is true for inappropriate gender forms.

The language pairs used for the experiments in this work are English–Russian (EN-RU) and English–French (EN-FR). For EN-RU, I use the parallel corpus prepared by Voita et al. [2019a][2] and also employed in Voita et al. [2019b]. This corpus consists of a training set of 6m sentence pairs, and a validation set and test set of 10k sentence pairs each. It was created by filtering the English–Russian OpenSubtitles2018 corpus for sentence pairs with a relative time overlap of subtitle frames of at least 90% to reduce noise in the data, and then randomly selecting 6.02m samples from this filtered corpus.

For EN-FR, I create a parallel corpus from the English–French OpenSubtitles2018 corpus of the same size as the EN-RU corpus. I first re-

[1] http://www.opensubtitles.org/

[2] https://github.com/lena-voita/good-translation-wrong-in-context#training-data

move duplicate samples from the corpus and then use Bicleaner [Sánchez-Cartagena et al., 2018] to remove noisy sentence pairs from the corpus. Bicleaner estimates the likelihood of a pair of sentences of being mutual translations based on a random-forest model. I apply the pre-trained EN-FR model[3] and sort the corpus by the likelihood score in descending order. From the sorted corpus, I take the top 6.02m parallel sentences, shuffle the corpus and tokenize it with the Moses tokenizer [Koehn et al., 2007]. Then I split it into a training set (6m), a validation set (10k), and a test set (10k).

Initially, I intended to append to both the EN-RU and the EN-FR corpora data from the Tatoeba[4] corpus. Tatoeba is a promising community project that aims to collect sentences and their translation for a wide range of languages through voluntary contributions. Often, several translation variants have been created for the same source sentence, differing, for example, in word choice, syntax, or features that are important to the target language but absent in the source language, such as politeness distinctions or gender variants. The availability of multiple translation variants for a single source sentence is an advantage over naturally grown corpora such as OpenSubtitles and would be ideal for training a gender-aware machine translation model.

Unfortunately, I found that the prevalence of male forms is even more pronounced in the Tatoeba corpus than in the OpenSubtitles corpus. For example, while in the EN-RU corpus created on the basis of OpenSubtitles2018 at least about 37.5% of all detected gender forms are female (see Section 4.1.2), the percentage of female forms in the EN-RU Tatoeba corpus is as low as 17.9%. This is an interesting finding in itself, as it shows that the masculine forms are still perceived as the standard form and the feminine forms as additional variants that are sometimes included, but often omitted. In any case, contrary to initial expectations, the addition of data from Tatoeba would exacerbate the pre-existing gender bias in the training corpus, so I refrain from using it for model training.

As part of my evaluation of gender-aware machine translation, I also conduct an experiment on triangular translation (see Chapter 5) from Russian via English to French (RU>EN>FR). The RU>EN and EN>FR models involved in this experiment are trained on the same data as described above.

[3] https://github.com/bitextor/bicleaner-data/releases/tag/v1.4
[4] https://tatoeba.org/

However, for evaluation purposes, I need an additional RU-FR test set. Unfortunately, Bicleaner only offers pre-trained corpus cleaning models for language pairs involving English as one of the languages, and the raw OpenSubtitles corpus is too noisy to use without cleaning. Therefore, I use the Russian–French sub-corpus of Tatoeba instead to create a test set.

The fact that I only need 10k samples out of a total of 245 077 parallel sentences in the RU-FR Tatoeba corpus allows me to be selective and create a test set with an approximately balanced gender ratio despite the gender imbalance in the Tatoeba corpus.

For sample selection, I apply the same script I use to annotate the training data with gender and politeness tags (see Section 4.1.2), and then randomly select 700 samples that contain a male gender reference, 700 samples that contain a female gender reference, 2000 samples that contain a politeness indication, and 6600 samples that do not contain any of the above according to the annotation script.

These numbers were selected such that they roughly correspond to the frequency of tags in the OpenSubtitles corpus (see section 4.1.2), except that the number of male and female references is more balanced. Due to the possible occurrence of both a female and a male reference in the same sentence, the final test set still has a slightly higher number of male references than female references (863 vs. 724). The test set is again tokenized with the Moses tokenizer.

4.1.2 Annotation of gender and politeness values

For the annotation of the EN-RU and EN-FR training corpora with gender and politeness tags, I develop a language-specific rule-based automatic annotation script for Russian and French. In this work, I use three tag sets, each with two mutually exclusive values:

tag	meaning	tag	meaning
<1m>	speaker is male	**<1f>**	speaker is female
<2m>	addressee is male	**<2f>**	addressee is female
<T>	form of address is informal	**<V>**	form of address is formal

Table 4.1: Tag set used in this thesis for both Russian and French

Tags are annotated only if a gender or politeness specific word form occurs in the sentence. For example, if a sentence contains no reference to the speaker or only a reference to the speaker in which no gender distinction is expressed, no tag from the corresponding tag set (1st person gender) is annotated. Therefore, a single sample can have between zero and up to three tags (one from each row in table 4.1).

All tags are annotated at the sentence level; in practice, they are prepended to the English source sentence. No word-level annotation is required because all relevant referents (e.g., the speaker) are unique to an entire sentence. The approach is intended to be general, and if desired, additional tag sets can be added to the tag collection.

For example, the English 2nd person pronoun *you* has the peculiarity that it is used for both singular and plural reference. This means that when translating a sentence that contains a reference to the second person into most other languages, it would also be necessary to specify the number (singular, plural) with an additional tag set (e.g., <2sg>, <2pl>). In Russian, however, 2nd person plural forms and the formal level of address are identical in most cases, and because of this overlap it is usually possible to create 2nd person plural references using the <V> tag. Therefore, an additional tag set is not needed for the EN-RU language pair.

The situation is more complex in the case of French. The plural forms of the 2nd person and the formal level of address do not coincide, so a fourth set of tags indicating the number of the 2nd person is necessary in principle so as to be able to produce all possible forms. However, in French, the masculine plural forms are used for mixed groups containing both males and females, while the feminine plural forms are used only when the group consists exclusively of females. Since a group of people (e.g., teachers) is most often mixed nowadays, feminine plural forms are rare. Because of this data sparsity, and to keep the tag collection consistent across the two languages used in this work, I refrain from adding this additional tag set, but note that it could be easily added if needed.

For corpus annotation, I make use of the morphological analyzers of spaCy [Honnibal and Montani, 2017]. Its Russian model comes from Nerus[5], and its French model comes from UD_French-Sequoia[6]. Initially, I intended to rely solely on these morphological analyzers for annotation

[5] https://github.com/natasha/nerus

[6] https://github.com/UniversalDependencies/UD_French-Sequoia

rules. Since gender and politeness distinctions occur only in certain morphological forms (e.g., gender in Russian past-tense verbs), resorting to a morphological parser is in theory the most principled way to automatically annotate the training corpus with gender and politeness values. Moreover, the method would allow relatively straightforward application to new languages for which a morphological parser is available.

However, although spaCy reports an accuracy of 97% of the morphological analyzer for Russian[7] and of 95% for French[8], I found that the models are not reliable enough for the purpose for which I intended to use them. In particular, imperative verb forms are frequently misclassified. Presumably, the morphology analyzers were trained on texts in which the first and second person occur infrequently, such as news texts or Wikipedia.

Therefore, I instead adopt a hybrid approach that takes into account both morphological information and the surface form of words in order to make a decision. To this end, I use the morphological analyzers to extract information about part of speech, gender, number, mood, and, as far as Russian is concerned, case. If applicable, this information is appended to the token with an underscore. For example, French *je me suis réveillée* (*I woke up*) is transformed into *je me suis réveillée_VERB_F*.

I then apply a hierarchical sequence of regular expressions to determine gender or politeness values based on the morphology-enriched text. If a match is found, the corresponding gender or politeness value is returned, otherwise the next regular expression is applied. The emphasis of the script is on high annotation accuracy rather than simplicity of rules. In any case, the number of regular expressions is too large to discuss them one by one. Instead, I provide here a general overview of the relevant phenomena in Russian and French. The full annotation scripts for Russian and French can be found on GitHub.[9]

Table 4.2 shows the main constructions where gender marking occurs in Russian with a reference to the speaker or the addressee. In general, gender agreement in Russian is relevant to verbs, adjectives and nouns. In reference to the speaker or addressee, gendered forms occur almost exclusively in predicative position, and in these cases the speaker or addressee is the subject of the sentence. The only exception in Table 4.2 are

[7] https://spacy.io/models/ru
[8] https://spacy.io/models/fr
[9] https://github.com/besou/genderMT

construction	example	translation
verbs in past tense	*Я **написал/а** письмо.*	*I **wrote** a letter.*
verbs in conditional mood	*Я **бы спросил/а** его.*	*I**'d ask** him.*
predicative adjectives	*Я **готов/а**.*	*I'm **ready**.*
predicative adjectives with verbs	*Я чувствую себя **счастливым/ой**.*	*I feel **happy**.*
predicative nouns	*Я **студент/ка**.*	*I'm a **student**.*
words of address (only 2nd)	*Привет, **дорогой/ая**!*	*Hi, **darling**!*

Table 4.2: Main phenomena involving gender marking in Russian.

construction	example	translation
verbs taking *être* in compound tenses	*Je suis **arrivé/e** à Moscou.*	*I **arrived** in Moscow.*
participles after direct object pronoun	*Il m'a **appelé/e** hier.*	*He **called** me yesterday.*
predicative adjectives	*Je suis **prêt/e**.*	*I'm **ready**.*
predicative nouns	*Je suis **étudiant/e**.*	*I'm a **student**.*
complements of transitive verbs	*Ça me rend **fou/folle**.*	*This drives me **crazy**.*
words of address (only 2nd)	*Salut, **chéri/e** !*	*Hi, **darling**!*

Table 4.3: Main phenomena involving gender marking in French.

words of address. However, the speaker or addressee being the subject of a sentence does not imply that their gender is necessarily expressed. For example, in *Я пишу письмо* (*I'm writing a letter*) no gender is expressed, while the same sentence in the past tense expresses the gender of the speaker (see Table 4.2).

Table 4.3 shows the most important constructions in which gender is marked in French with a reference to the speaker or the addressee. The phenomena in French are similar to those in Russian, albeit not identical. The main difference is with verbs. Unlike in Russian, not all past-tense verbs convey gender, but instead all compound forms of verbs that take *être* (*to be*) as their auxiliary. This is the case for a relatively small number of mostly movement-related verbs, such as *venir* (*to come*), and for all reflexive verbs. Verbs that use *avoir* (*to have*) as an auxiliary verb do not in general inflect for gender. A peculiarity of French, however, is that compound forms of verbs taking *avoir* convey gender if (and only if) the direct object precedes the participle, for example in *Il m'a appelée hier* (*He called me* [female] *yesterday*). Unlike in Russian, gender agreement for the 1st and 2nd person can therefore occur both when the 1st or 2nd person is the subject and when it is the direct object.

construction	example	translation
2nd person personal pronouns (*you*)	*Могу я пойти с **тобой/вами**?*	*Can I come with **you**?*
2nd person possessive pronouns (*your*)	*Это **твоя/ваша** сумка?*	*Is this **your** bag?*
verbs in the imperative	***Послушай/те** меня!*	***Listen** to me!*

Table 4.4: Main phenomena involving TV marking in Russian.

construction	example	translation
2nd person personal pronouns (*you*)	*Je peux venir avec **toi/vous** ?*	*Can I come with **you**?*
2nd person possessive pronouns (*your*)	*C'est **ton/votre** sac ?*	*Is this **your** bag?*
verbs in the imperative	***Écoute/z**-moi !*	***Listen** to me!*

Table 4.5: Main phenomena involving TV marking in French.

As far as politeness distinctions are concerned, Russian and French are very similar. Both distinguish two levels of politeness, which may be called *formal* and *informal*, and these are expressed when a reference is made to the addressee (2nd person). Tables 4.4 and 4.5 show the relevant constructions in which a politeness level is expressed. Unlike gender marking, politeness marking is not limited to sentences in which the addressee is the subject. Also in contrast to gender, politeness is expressed mainly in pronouns. These include the personal pronouns (*you*) in all case forms and the possessive pronouns (*your*) in all gender, number and, in Russian, case forms. Furthermore, all 2nd person verb forms express a politeness level. However, the finite verb is almost always accompanied by a personal pronoun that already expresses the same politeness level and, therefore, the marking of the verb is redundant; hence, this phenomenon is not listed in the tables. The only exception are verbs in the imperative, which are regularly used without a subject pronoun.

Although the phenomena of gender and politeness in French and Russian are relatively simple in theory, it is not entirely trivial to capture all the variants of these phenomena automatically. The concrete difficulties differ slightly between Russian and French. Russian has a relatively unambiguous morphology but a free word order, which can present pitfalls for rule-based pattern matching. French, on the other hand, has a fixed word order, but the morphology is much more opaque, so that it is often impossible to deduce the gender of a word in a regular way from its surface form.

tag	EN-RU		EN-FR	
	# samples	% samples	# samples	% samples
1st or 2nd person	736 532	12.28	428 438	7.14
male	467 138	7.79	299 395	4.99
female	279 817	4.66	131 236	2.19
1st person	471 634	7.86	250 460	4.17
male	295 046	4.92	177 005	2.95
female	176 588	2.94	73 455	1.22
2nd person	286 664	4.78	182 631	3.04
male	179 793	3.00	124 221	2.07
female	106 871	1.78	58 410	0.97
TV distinction	1 750 317	29.17	1 772 878	29.56
informal	1 172 661	19.54	816 968	13.62
formal	577 656	9.63	955 910	15.93

Table 4.6: Number and percentage of samples containing tags per tag set in the annotated EN-RU and EN-FR training corpora.

Table 4.6 shows the number and percentage of samples in the EN-RU and EN-FR training corpora that exhibit a particular tag, as annotated by the rule-based scripts. The number of samples containing a politeness distinction is very similar in both languages, with 29.17% in Russian and 29.56% in French. These percentages also correspond to the numbers given in Sennrich et al. [2016a] for politeness expressions in German (1.57m out of 5.58m samples = 28.14%). The numbers for formal and informal politeness levels also reveal that French has a stronger preference for formal forms of address (15.93%) compared to Russian (9.63%).

Gender forms referring to the 1st or 2nd person are less frequent, at 12.28% in Russian and 7.14% in French. The large percentage difference between the two languages results from the fact that gender does not occur in the same constructions in the two languages. In Russian, the most common phenomenon of gender agreement with the 1st or 2nd person are verbs in the past tense. Most past-tense verbs in French, on the other hand, do not express gender.

However, in both corpora feminine forms are underrepresented. The proportion of samples with female references out of all samples containing

verb	translation	masculine		feminine	
был/а	*was*	92 413	61.9 %	56 884	38.1 %
сказал/а	*said*	37 177	66.3 %	18 930	33.7 %
хотел/а	*wanted*	23 259	62.2 %	14 160	37.8 %
думал/а	*thought*	14 646	60.2 %	9 693	39.8 %
мог/ла	*was able*	19 287	67.0 %	9 484	33.0 %
сделал/а	*did*	19 237	70.2 %	8 177	29.8 %
видел/а	*saw*	14 897	66.0 %	7 683	34.0 %
знал/а	*knew*	13 866	63.3 %	8 025	36.7 %
говорил/а	*talked*	13 928	64.7 %	7 600	35.3 %
нашёл/шла	*found*	8 117	63.8 %	4 605	36.2 %

Table 4.7: 10 most common Russian past tense verbs in the EN-RU training corpus and their counts and share per gender.

tag	EN-RU		EN-FR	
	# samples	% samples	# samples	% samples
only 1st person	338 772	5.65	184 444	3.07
only 2nd person	10 673	0.18	15 207	0.25
only politeness	1 363 596	22.73	1 544 978	25.75
1st person + 2nd person	366	0.00	887	0.01
1st person + politeness	111 096	1.85	61 363	1.02
2nd person + politeness	254 225	4.24	162 771	2.71
all three	21 400	0.36	3 766	0.06
total	2 100 128	35.00	1 973 416	32.89

Table 4.8: Number and percentage of samples containing specific tag combinations in the annotated EN-RU and EN-FR training corpora.

a gender reference is 37.99% in Russian and 30.63% in French. In other words: In about one third of the sentences the speaker or the addressee is a woman, while in two thirds of the sentences it is a man.

To make sure that this gender imbalance in the annotated corpus is not due to a bias in my annotation scripts, I double-check it with a simple statistic. For this, I extract the 10 most frequent verbs in the past tense from the EN-RU corpus and compare their number per gender. Since past tense verbs in Russian do not have person-specific morphology, these counts include all verbs with 1st, 2nd and 3rd person singular subjects.

The counts for these verbs show a similar pattern, with the proportion of feminine forms ranging from 29.9% to 39.8%. It can therefore be assumed that the annotation scripts themselves do not have a gender bias, but that the number of tags reflects the actual distribution in the data quite closely.

For reference, Table 4.8 reports the number of samples with specific combinations of tags, such as a 2nd person tag and a politeness tag. The strongest conclusion that can be drawn from this table is that when the gender of the 2nd person is expressed, the level of politeness is usually expressed as well. Therefore, 2nd person gender tags are rare unless coupled with a politeness tag.

4.1.3 Subword segmentation

Segmenting words into subwords is a successful method to solve the problem of out-of-vocabulary tokens. It also helps to keep the vocabulary size small. In this work, I use the unigram subword segmentation model [Kudo, 2018] implemented in the *sentencepiece* library [Kudo and Richardson, 2018].

Different from Byte Pair Encoding (BPE) [Sennrich et al., 2016b], another well-established segmentation algorithm, the unigram model is a probabilistic model, i.e., it assigns a probability to each possible segmentation of a word. In its simplest form, it divides a word into subwords such that the total probability of the sequence is maximal. However, it is also possible to sample different segmentations for the same word based on the probabilities. Employing different word segmentations for the same word (which is also called subword regularization) renders the translation model more stable against spelling errors, among other things.

The probability of a word segmentation is calculated as the product of the independent probabilities of the individual segments, hence the name *unigram* model. The most probable segmentation of a word is found with the Viterbi algorithm [Viterbi, 1967]. As with BPE, subword segmentation is applied only within words, that is, the segments do not cross word boundaries.

The unigram word segmentation model is trained as follows. First, a large seed vocabulary is initialized heuristically based on the training corpus. This vocabulary may, for example, consist of all characters and the most frequent substrings in the corpus. Then the following three steps are repeated until the desired vocabulary size V (a hyper-parameter) is

reached: 1) fixing the vocabulary, optimize all subword occurrence probabilities with the EM (expectation maximization) algorithm; 2) calculate the loss for each subword, where loss is defined as the reduction in probability if that subword were removed from the current vocabulary; 3) sort the subwords by loss and remove the bottom η% (e.g., 20%). Subwords consisting of single characters are never removed to ensure that any token can be segmented. The intuition behind this algorithm is that the original vocabulary is reduced to the desired size by removing the rarest subwords. Since the real probability of a subword is unknown, it is estimated iteratively.

In this work, I train a unigram model for each language with a vocabulary size of 32k subwords each. I apply the model to the corpus using one-best decoding, i.e., each word is consistently segmented into the sequence with the highest probability.

4.2 Model

4.2.1 Architecture

I employ standard transformer models as described in Vaswani et al. [2017] implemented in the *fairseq* toolkit [Ott et al., 2019] for my experiments. The transformer architecture is a variant of an encoder-decoder architecture and as such consists of a module that encodes the source sequence into an internal representation (the encoder) and a module that decodes the internal representation into the target sequence (the decoder). The encoder-decoder architecture is useful for sequence-to-sequence prediction tasks where the length of the source and target sequences do not necessarily match, as is the case with machine translation.

In the case of the transformer, the encoder consists of *N* identical layers stacked on top of each other. Each layer consists of two sub-layers: multi-headed self-attention (see below), and a position-wise fully connected feed-forward network. Each sub-layer has a residual connection around it, i.e. the output of each sub-layer is summed with the input. In addition, a layer normalization [Ba et al., 2016] is applied to the raw output vector of each sub-layer.

The architecture of the decoder is very similar to that of the encoder. It is also composed of *N* identical layers stacked on top of each other. Each

layer consists of three sub-layers: masked multi-headed self-attention, followed by multi-headed attention (to the encoder), followed in turn by a fully connected feed-forward network. Again, each sub-layer has a residual connection around it, and its output is layer-normalized. The output of the *N*-th decoder layer is then passed to a linear layer with softmax, which produces the output probabilities over the vocabulary for each position.

The key component of the transformer architecture is the attention mechanism, which is used in both the encoder and decoder parts of the model. Conceptually, attention allows each token to see all tokens in a sequence, regardless of their position, and to refine its own encoding based on this information. In the self-attention module of the encoder, each token can look at all other tokens in the same sequence and thereby identify other tokens that might be relevant to its own encoding. For example, a finite verb may attend to its subject and/or object and encode this information in its own representation.

The self-attention module of the decoder is similar to that of the encoder, except that each position can only attend to previous positions, while the positions to the right are masked. This is necessary due to the autoregressive (left-to-right) generation of the output sequence. Thus, each position can attend to previously generated words in the output sequence and take this information into account for selecting the next output token.

Finally, the (decoder-to-encoder) attention module in the decoder allows each decoder position to attend to all tokens in the encoded source sequence, thereby identifying relevant tokens at the current position. This can be, for example, the word in the source sentence for which a translation is generated at the current position.

In technical terms, the attention mechanism employed in Vaswani et al. [2017] can be described using terminology of database retrieval. Three vectors are used to calculate an attention value: a query, a key, and a value. The query, key and value vectors are linear transformations of vectors of word tokens. The matrices used for this linear transformation are learned during training. The attention value is calculated by taking the dot product of the query and the key, and then multiplying the resulting scalar by the value vector. In reality, all attention values for a specific query are calculated in parallel, using matrices instead of vectors in the input and output. In addition, the values in the intermediate vector (containing the results of the dot products) are first scaled by the dimension of key and then softmax is applied to the vector, before multiplying it with the value matrix.

The scaling is done to avoid very small gradients in the softmax function (see Vaswani et al. [2017] for more details).

The attention mechanism is applied multiple times in the same way (but with potentially different transformation matrices) for each token. This is called multi-head attention and allows the different 'heads' to focus on different aspects. For example, one attention head may learn to focus on syntactic information and another to focus on word sense disambiguation.

The final sub-layer of each layer in the encoder and decoder is a position-wise feed-forward network. This is a standard two-layer feed-forward neural network with a ReLU activation function that is applied to each position individually. The input and output vectors have a dimension of 512, and the inner layer has a dimension of 2048. The feed-forward network serves to process the information obtained by the preceding attention mechanism.

An important innovation of the transformer is that it does not process the source tokens one after another, but all of them in parallel. In order for the model to still make use of the order of sequence, an encoding of the absolute or relative position is needed. Vaswani et al. [2017] use sine and cosine functions of different frequencies for this positional encoding. The positional encodings have the same dimension as the input embeddings. They are summed before being passed to the encoder.

In my experiments, I use the transformer architecture with the same hyper-parameters as in Vaswani et al. [2017]. The encoder has 6 layers, an embedding dimension of 512, a feed-forward embedding dimension of 2048 and 8 attention heads. ReLU is used as the activation function. The decoder has the same hyper-parameters as the encoder.

4.2.2 Training

The models are trained for 25 epochs (approximately 160k updates). Cross entropy is used as the loss function and Adam [Kingma and Ba, 2015] with parameters β_1 = 0.9 and β_2 = 0.98 as the optimizer. The initial learning rate is set to 5e-4 and the learning rate schedule is the inverse square root schedule as used in Vaswani et al. [2017]. In this schedule, the learning rate is increased linearly for the first 4 000 training steps and then decreased proportionally to the inverse square root of the step number.

Dropout with rate = 0.3 [Srivastava et al., 2014] is used as a regularization technique. It is applied to the raw output of each sub-layer (before its summation with the residual vector), as well as to the sum of the embeddings and the positional encodings. As another regularization technique to avoid overfitting, early stopping is activated, however, all models continued to train until epoch 25. Finally, label smoothing [Szegedy et al., 2016] with parameter = 0.1 is employed to prevent overconfidence of the model.

In each epoch, training samples are shuffled and training samples of similar length are batched together. The maximal number of tokens per batch is 4 096. Beam search with a beam size of 5 is used for generation.

In order for the model to learn to both deal with missing tags and ignore irrelevant tags, some of the gold standard tags are removed and others are added randomly during training. I follow the approach described in Sennrich et al. [2016a] and re-tag the training corpus for each epoch. In the paper mentioned, this is done by randomly removing 50% of the correct tags and adding a random tag to 50% of the samples that do not have a tag. This prevents the model from learning that the presence of a marker is in itself significant.

However, in contrast to the paper mentioned above, where only a single tag set is used for the politeness level (<T> vs. <V> in the terminology applied here), there are three different binary tag sets in this work, one for the gender of the speaker (<1m> vs. <1f>), one for the gender of the addressee (<2m> vs. <2f>), and one for the politeness level (<T> vs. <V>). Since the model should learn that tags from different tag sets can be applied independently of each other, I also need to prepare the tag sets independently from each other. Therefore, for each epoch, I randomly remove 50% of the correct tags *per tag set*, and accordingly add a random tag to 50% of the samples that do not have a tag *from the respective tag set*. This ensures that the model is prevented from interpreting the mere presence of a tag, as discussed above. However, unlike in the case with a single tag set, this approach results in a corpus in which the majority of all samples (approximately 7/8) has at least one tag. This is because each sample has a probability of 50% of having a tag for the 1st person gender, but also the same probability of having a tag for the 2nd person gender, and again the same probability of having a tag for the politeness level. This means that during training, the model is exposed to far more samples with a tag — be it relevant or not — than to samples without a tag. The effects of this will be further discussed in Chapter 6.

5 Experiments

5.1 Oracle tagging

In the first and main experiment, I evaluate whether passing appropriate gender and politeness tags along with the input improves translation quality in terms of BLEU. In addition, I investigate whether the tags are correctly interpreted and accounted for in the output. To this end, I compare the performance of the gender-aware translation model on correctly tagged samples to the performance of a baseline model on samples without tags.

From a practical point of view, the experiment simulates a situation where a user passes tags along with the input sentence to control the gender and politeness values in the generated translation, such that the resulting translation then adequately reflects the user's specifications.

In automatic evaluation using a parallel test corpus, there is no user to specify the values for gender and politeness. Instead, we have a reference translation that already contains fixed values for gender and politeness that the generated translation must match. Hence, we can approximate the real use case by inferring gender and politeness values from the reference translation.

This type of experiment, where the model has access to information that it does not have access to in practice (in this case, metadata about the reference translation), is sometimes called an oracle experiment in natural language processing. A comparable experiment was conducted in Sennrich et al. [2016a] to evaluate the usefulness of politeness tags.

The two models EN>RU and EN>FR are evaluated on the respective test set splits with 10k samples each (see Chapter 4). The samples are tagged with the same scripts that were used to tag the training corpora (see Chapter 4). To assess the impact of tagging in general, the untagged samples

are translated once with the baseline models and once with the gender-aware models.

BLEU scores are calculated using sacreBLEU [Post, 2018] on mixed-case text, using the 13a tokenizer and an n-gram width of 4. They are reported for the entire test set as well as for various subsets, e.g., for samples containing masculine forms, for samples containing feminine forms, and for samples containing politeness forms in the reference translation.

BLEU scores can be seen as a proxy for evaluating the correctness of gender and politeness forms: if the gender and politeness forms are correct in the translation, more words and more word n-grams will be correct overall, and thus the BLEU score will be higher. However, BLEU does not explicitly evaluate whether gender and politeness tags are indeed reflected accurately in the translation.

Therefore, I additionally apply the rule-based tagging script to the translation hypotheses and report for each subset of the test set the number of matches and mismatches with the tags present in the input. The agreement between the detected gender or politeness values in the output and the input tags is a measure of how reliably the model accounts for gender and politeness constraints.

Finally, in order to analyze effects of a possible gender bias in the model, I evaluate translation quality on a test set of 200 EN-RU examples for which I manually created a reference translation in the other gender. I evaluate both BLEU and gender accuracy on four variants of this test set: 1) the original samples as attested in the corpus; 2) the samples with reversed gender forms in the reference translation; 3) the samples with exclusively male gender forms in the reference translation; 4) the samples with exclusively female gender forms in the reference translation.

The original test set consists of 200 randomly selected samples from the full EN-RU test set that contain at least one gender tag. Samples that contain explicit gender references in the original English, such as *You're a good man, Daniel*, were filtered out manually.

In contrast to the full test set, this parallel test set allows to evaluate the performance of the gender-aware model on sentences that differ only in the attributes of gender and politeness and are otherwise equivalent. Table 5.1 shows one sample from the test set in masculine and feminine configuration.

input	*Listen, I **was** over at the Cardozo Farms the other day.*
reference, male	*Слушай, я недавно **был** на ферме Кардозо.*
reference, female	*Слушай, я недавно **была** на ферме Кардозо.*

Table 5.1: Example of two equivalent reference translations differing only in gender forms.

5.2 Random tagging

In the second experiment, I investigate how random tagging affects translation quality. Sometimes users may not want to or are unable to specify gender or politeness tags. In this case, the model can simply be used without tags. Perhaps surprisingly, however, enriching the input with random tags offers at least two interesting use cases, which I call balanced random tagging and consistent random tagging. Both use cases do not require user interaction.

First, machine translation models often exhibit a strong gender bias, either applying gender forms based on stereotypes or favoring one gender, usually male, in most contexts. This issue is highly prevalent even in the most advanced commercial systems such as DeepL or Google Translate, as shown in Chapter 2. Gender-aware machine translation models have the potential to provide a simple solution to this problem: applying female and male tags randomly with equal probability. If the tagging approach works, this should considerably reduce gender bias in the output, even if the underlying translation model exhibits a gender bias.

Second, a sentence-level machine translation model cannot guarantee a consistent politeness level or gender for a given person across sentences, which can be confusing for readers. If the gender of the speaker, the gender of the addressee or the social situation is unknown, either level of politeness and either gender can be fine in translation, but the choice should be consistent throughout the text. Again, if the tagging approach works, a gender-aware machine translation model can be used to consistently apply the same random tag configuration to all sentences in an entire text. This is expected to increase consistency of gender and politeness values across a translated text.

For the purpose of examining random tagging, I use the same test sets for EN-RU and EN-FR as above, each with 10k samples, and evaluate both BLEU scores and gender accuracy for four differently tagged variants: 1)

the raw input with no tags; 2) male tags for both persons; 3) female tags for both persons; 4) half of the samples with male tags and the other half with female tags.

Random tagging is not expected to improve translation quality in terms of BLEU; instead, performance should ideally be relatively stable across different tagging variants. More important is the analysis of matches and mismatches between the gender tags in the input and the detected gender in the output, which quantifies the extent to which the constraints are taken into account by the model. Special attention is also given to cases where the original text contains an explicit gender reference and the input tags are in contradiction to it, such as, e.g., *You're my brother* combined with a tag indicating a female addressee.

5.3 Tag prediction in triangular translation

English is often used as a pivot language in triangular translation. In triangular translation, a sentence is not translated directly from language A into language B, but the original sentence in language A is first translated into English (or another pivot language) and then translated from English into language B.

Triangular translation can be beneficial, for example, if there is not enough parallel data available for languages A and B. However, indirect translation can also lead to a deterioration in translation quality due to error propagation and, equally important, information loss in the pivot language.

For example, when translating from Russian into French via English, the gender information contained in the Russian sentence is lost in English and therefore cannot be transferred into French:

RU		EN		FR
Я счастлив.	$\rightarrow$	*I'm happy.*	$\overset{?}{\rightarrow}$	*Je suis heureux.*
Я счастлива.			$\overset{?}{\rightarrow}$	*Je suis heureuse.*

However, we can avoid this loss of information in English by enriching the English sentence with appropriate tags that may be automatically derived from the original Russian sentence:

RU		EN		FR
Я счастлив.	→	<1m> *I'm happy.*	→	*Je suis heureux.*
Я счастлива.	→	<1f> *I'm happy.*	→	*Je suis heureuse.*

In this third experiment, I investigate the usefulness of tag prediction in triangular translation using Russian-English-French (RU>EN>FR) as an example. This approach requires not only an English–French model which can interpret tags (as in the first two experiments), but in addition, a Russian–English model which can output appropriate tags together with the English translation. For this, I train a model on the same English–Russian data, including the tags, and with the same hyperparameters, while simply reversing the source and target languages.

I evaluate the triangular translation model RU>EN>FR on a Russian-French (RU-FR) test set with 10k samples created from the Russian-French Tatoeba corpus (see Chapter 4 for more details). As in the first two experiments, I report BLEU values for the whole test set and for different subsets of the test set, e.g. for samples containing masculine forms or samples containing feminine forms.

For the baseline model, I remove all tags predicted by the RU>EN model before passing the English sequences to the EN>FR model. This approach ensures that the baseline model and the gender-aware model differ only in the tags, which makes it possible to measure the effects of the tags themselves and not of other possible differences in triangular translation that would arise if a tag-free RU>EN model were used instead.

A special case of triangular translation is back-translation. Here, the original sequence is translated into another language and back into the original language, e.g. from Russian into English and back into Russian. I also conduct an experiment on back-translation into Russian (RU>EN>RU) using the same approach as described above. Since the source and target languages are identical in back-translation, this method is particularly well suited to quantifying the reduction in information loss made possible by the tagging approach. The evaluation is carried out on the EN-RU test set that was also used in the previous experiments.

In triangular translation, the tags for gender and politeness are only used internally and do not need to be passed on to the model. This allows a direct comparison of the tagging approach with commercial translation

systems. Therefore, I also compare the performance of the RU>EN>FR model with the translations of DeepL from Russian to French.

6 Results

6.1 Oracle tagging

6.1.1 Evaluation with BLEU

Table 6.1 reports the results for the English–Russian (EN>RU) models on the full test set of 10k samples and on different subsets of the test set. Results are given for three configurations: The first column (*baseline*) contains the BLEU scores for the baseline model trained on the untagged dataset. The second column (*w/o tags*) contains the BLEU scores for the model trained on the tagged dataset but used without tags during testing. Finally, the third column (*w/ tags*) contains the BLEU scores for the same model, and this time with the correct tags prepended to the input. For brevity, I refer to this last model in what follows as the *gender-aware* model. The last column of the table shows the improvement of the gender-aware model over the baseline model.

I report values for various subsets of the test set. The upper section of the table includes the values for: 1) all samples in the test set (*full test set*); 2) all samples that do not contain words in the reference translation that indicate the gender of the speaker, the gender of the addressee, or a politeness level (*cont. no gender/TV*); 3) all samples that exhibit at least one of the three phenomena (*cont. gender/TV*).

The samples containing words indicating 1st or 2nd person gender and/or a politeness level are further analyzed in the subsequent two sections of the table. The middle section focuses on samples that contain words expressing 1st or 2nd person gender and gives the values for: 4) all samples that contain words in the reference translation that indicate the gender of the speaker or recipient (and possibly words that express a politeness level in addition) (*cont. gender*); 5) all samples containing words

test set	# samples	baseline	w/o tags	w/ tags	improvement
full test set	10 000	30.74	30.47	**33.08**	+2.34
cont. no gender/TV	6 498	30.26	30.76	**30.76**	+0.50
cont. gender/TV	3 502	31.33	30.10	**35.97**	+4.64
cont. gender	1 262	32.06	28.18	**37.73**	+5.67
cont. masculine	784	33.82	28.77	**36.88**	+3.06
cont. feminine	499	29.56	26.92	**39.53**	+9.97
cont. only TV	2 240	30.84	31.36	**34.78**	+3.94
cont. T	1 309	35.63	35.53	**37.44**	+1.81
cont. V	931	24.58	25.94	**31.38**	+6.80

Table 6.1: BLEU scores for EN>RU on various subsets of the test set. Best values in bold. The last column shows the improvement of oracle tagging (w/ tags) over the baseline model.

referring to a male speaker or addressee (*cont. masculine*); 6) all samples containing words referring to a female speaker or addressee (*cont. feminine*).

The samples containing words indicating politeness level (but no words indicating gender) are analyzed in the lower part of the table. All these samples have exactly one tag, either <T> or <V>. Values are reported for: 7) all samples containing words indicating a politeness level (*cont. TV*); 8) all samples containing words indicating an informal level of politeness (*cont. T*); 9) all samples containing words indicating a formal level of politeness (*cont. V*).

Note that the sum of samples with masculine forms (784) and samples with feminine forms (499) is greater than the number of samples with gendered forms (1 262). This is because some samples express both the gender of the speaker and the gender of the addressee, and if the gender is not the same for both parties, these samples appear once in each subset. One such example is *Je suis navrée, mon chéri* (*Honey, I'm sorry*), where the speaker is female and the addressee is male.

The results are encouraging and very consistent: The gender-aware model exceeds the baseline model on all subsets, and the improvement is statistically highly significant ($p < 0.01$) on the full test set and on all subsets containing gender or politeness expressions.[1] On the full test set

[1] All significance tests are performed with SacreBLEU using paired bootstrap resampling.

containing a random selection of samples, tagging improves the model by +2.34 BLEU points. On the subset of samples with a gender or politeness expressions, the improvement is +4.64 BLEU points. The improvement is higher for samples with a gender expression (+5.67) than for samples with only a politeness expression (+3.94). It should be noted, however, that the former subset includes all samples with tag combinations.

With both gender and politeness level, the improvement is particularly high for the variant which is found less often in the training data and thus disfavored by the translation model. The improvement is as high as +6.80 BLEU for samples containing a polite form, and +9.97 BLEU for samples containing a feminine form. This indicates that tags are particularly useful to generate variants disfavored by the model.

Surprisingly, the BLEU score of the gender-aware model is higher for samples with feminine forms (39.53) than for samples with masculine forms (36.88), although the opposite is true for the baseline model (29.56 vs. 33.82). The same cannot be observed for politeness, as in both the baseline model and the gender-aware model, the BLEU score is higher for T-forms (35.63 and 37.44, respectively) than for V-forms (24.58 and 31.38, respectively).

The results for the model trained on tagged data but used without tags during testing (*w/o tags*) are inconclusive. Ideally, the performance should match that of the baseline model, since the input is the same. This is indeed the case when looking at the full test set, where it has a BLEU value of 30.47, which is close to the baseline model with 30.74. The same is true for samples containing politeness-related expressions, with a value of 31.36 compared to the baseline with 30.84. However, for samples containing gender-specific expressions, the BLEU values of 28.18 are significantly lower compared to the baseline model, which is at 32.06.

Table 6.2 reports the results for the English-French (EN>FR) model on the full test set of 10k samples and on various subsets of the test set. Similar to EN>RU, the results for EN>FR are consistent across all subsets: The gender-aware model outperforms the baseline model on all subsets, and the improvement is statistically highly significant ($p < 0.01$) on the full test set and on all subsets containing gender or politeness expressions.

On the full test set, tagging improves the model by +2.14 BLEU (vs. +2.34 BLEU for EN>RU). On the subset of samples with a gender or politeness expressions, the improvement is +5.57 BLEU (vs. +4.64 BLEU for EN>RU). While for EN>RU the improvement was more pronounced in the

test set	# samples	baseline	w/o tags	w/ tags	improvement
full test set	10 000	45.96	46.05	**48.10**	+2.14
cont. no gender/TV	6 725	47.00	47.25	**47.25**	+0.25
cont. gender/TV	3 275	44.07	43.85	**49.64**	+5.57
cont. gender	716	44.55	43.17	**50.16**	+5.61
cont. masculine	501	44.75	43.88	**49.17**	+4.42
cont. feminine	218	44.40	42.04	**52.83**	+8.43
cont. only TV	2 559	43.91	44.06	**49.46**	+5.55
cont. T	1 159	44.50	44.32	**50.65**	+6.15
cont. V	1 400	43.46	43.87	**48.55**	+5.09

Table 6.2: BLEU scores for EN>FR on various subsets of the test set. Best values in bold. The last column shows the improvement of oracle tagging (w/ tags) over the baseline model.

gender samples compared to the politeness samples, the improvement for EN>FR is about the same for both subgroups (+5.61 BLEU for cont. gender and +5.55 BLEU for cont. only TV).

The reason for this could be that unlike the EN-RU data, where T-forms are much more common than V-forms, T-forms and V-forms are about equally common in the EN-FR training corpus (see 4). Therefore, the baseline model for EN>FR cannot as easily achieve a good performance by generally favoring the more frequent variant over the other.

As with EN>RU, we observe the unexpected fact that in the gender-aware model, the BLEU score for samples with feminine forms (52.83) is considerably higher than for samples with masculine values (49.17), and in fact the BLEU score for samples with feminine forms is generally the highest observed across all subsets.

As for the model trained on tagged data but used without tags during testing (*w/o tags*), its BLEU values are very close to the baseline model, as would be expected. On the full test set, the BLEU value is 46.05 compared to 45.96 for the baseline model, and all subsets have values either slightly below or slightly above the baseline model.

The BLEU scores for the EN>FR models are generally much higher than those for the EN>RU models. For example, the gender-aware EN>FR model has a score of 48.10 BLEU, while the EN>RU model has a score of 33.08 BLEU. This is expected given the lexical and syntactic proximity of

English and French on the one hand, and the much greater structural difference between English and Russian on the other.

The results are very consistent across both language pairs. This is noteworthy because both the training set and the language-specific annotation rules described in Chapter 4 are different for each language. The largest difference between the two language pairs is observed in the samples containing only politeness expressions. This is understandable, because French and Russian do indeed differ in their use of the politeness expressions. Overall, formal expressions are used more extensively in French, as can be seen from the corpus statistics in Chapter 4.

As a final note, multiple tags in a sample always appear in a fixed order, both in the training sets and in the tagged test sets: first the gender of the speaker, then the gender of the addressee, and finally the politeness level. For example, this may look like this: '<1f> <2m> <T>'. To evaluate whether the model relies on this order, I evaluated the EN>RU model once again on the subset containing gender or politeness expressions, but with disturbed tag order. For all samples containing two tags, I simply swapped the two tags, and for samples containing three tags, I randomly rearranged the three tags. The BLEU score on this test set is 35.99 compared to 35.97 on the original test set. This difference is not statistically significant ($p > 0.05$), which indicates that the model has learnt that the order of the tags does not matter – even though it has been trained on tags in a fixed order.

6.1.2 Evaluation of gender correctness

model	test set	# tags	baseline				gender-aware			
			m	f	n/b	error rate	m	f	n/b	error rate
EN>RU	male	793	493	90	210	11.35%	634	1	158	0.13%
	female	505	286	100	119	56.63%	2	431	72	0.40%
EN>FR	male	506	250	34	222	6.72%	294	3	209	0.59%
	female	218	73	68	77	33.49%	1	148	69	0.46%

Table 6.3: Confusion matrix of gender in the reference translations and detected gender in the hypotheses (male, female, none/both) of the baseline model and the gender-aware model for EN>RU and EN>FR. The error rate is the percentage of opposite-gender forms out of the total number of tags.

While BLEU scores provide an indication of the usefulness of gender tags, they do not explicitly assess whether the tags are used by the model

model	type	text	gender
EN>RU	input	<2f> <T> *Do you wanna say that a little louder?*	f
	reference	*Ты не могла сказать это еще громче?*	f
	hypothesis	*Не хочешь сказать это погромче?*	n/b
EN>FR	input	<1m> *I was in Naples.*	m
	reference	*Je suis resté à Naples.*	m
	hypothesis	*J'étais à Naples.*	n/b

Table 6.4: Two examples of gender-specific forms in the reference translations and gender-unspecific forms in the automatic translations.

in the intended way. Therefore, as a second part of the evaluation, I use the rule-based annotation scripts to annotate the detected gender forms in the model hypotheses and compare them to those in the reference translations. Table 6.3 reports the results of this evaluation for both EN>RU and EN>FR.

As can be seen from the table, a relatively large number of hypotheses falls into the class *n/b*, i.e., they do not contain a gender-specific expression. This is somewhat unexpected, since all references contain gender-specific expressions. The reason for this is different syntactic constructions in the hypothesis that do not require gender marking and are typically caused by a more literal translation from English compared to the reference. Table 6.4 shows two examples of this phenomenon. For example, the English *I was* is translated in the reference as *Je suis resté*, which is only used for a male speaker, whereas the construction for a female speaker would be *Je suis restée*. On the other hand, the model translates this phrase into French as *J'étais*, which fits both genders. The gender-unmarked hypotheses can be considered correct because they do not contradict the desired gender.

It is evident that the baseline models have a bias towards the male gender. For example, out of the 505 tags in the EN-RU test set that refer to a female, only 100 are translated by the baseline model in the female form and 286 as male. This is not surprising; since the baseline model does not know the correct gender, it tends to guess the one that is more common in the training data. This corresponds to an error rate of 56.63%. The same is true for EN>FR, albeit to a lesser extent, with an error rate of 33.49% for female references.

model	type	text	gender
EN>RU	input	<1m> *Never felt more alive [...]*	m
	reference	*Я почувствовал себя таким живым [...]*	m
	hypothesis	*Никогда не чувствовала себя более живым [...]*	f
EN>FR	input	<1f> *Dad, been kidnapped.*	f
	reference	*Papa, j'ai été kidnappée.*	f
	hypothesis	*Papa, j'ai été kidnappé.*	m

Table 6.5: Two examples of incorrect gender forms in the model output. (Note: The Russian hypothesis actually contains word forms in both genders in a contradictory way.)

On the other hand, the gender-aware models almost always generate hypotheses with appropriate gender forms. They consistently have error rates below 0.6%. For example, for EN>RU and female references, oracle tagging reduces the error rate, i.e. the rate of undesired gender forms, from as high as 56.63% to only 0.40%. Due to the bias towards the male gender, the error rates of the baseline models for masculine forms are much lower than those for feminine forms. Still, the error rates for male-related samples are also greatly reduced with the gender-aware models, from 11.35% to 0.13% in the case of EN>RU and from 6.72% to 0.59% in the case of EN>FR. These results show that the models do indeed interpret the tags correctly and as expected.

Table 6.5 shows two of the total seven samples with incorrect gender forms in the output. Both examples have in common that the subject pronoun is missing, which I also observed as a difficulty when trying out the models interactively.

6.1.3 Evaluation on equivalent sentences

The test set used in the main evaluations above represents real data. Therefore, the sentences referring to a male person and those referring to a female person are completely different sentences and unrelated to each other. In this section, performance is instead evaluated on equivalent sentences to better understand the extent and nature of gender bias in the model itself.

The test set consists of 200 samples that I randomly selected from those samples in the EN-RU test set that contain gender references. For these sentences, I manually created a reference translation in the oppo-

test set	# samples	baseline	w/o tags	w/ tags	improvement
original	200	30.14	29.12	**37.52**	+7.38
gender-reversed	200	29.14	27.75	**36.11**	+6.97
all masculine	200	33.40	30.04	**37.17**	+3.77
all feminine	200	25.72	26.82	**36.58**	+10.86

Table 6.6: BLEU scores for EN>RU on 200 sentences that differ only in gender and are otherwise equivalent.

site gender. Four different configurations are evaluated: 1) the original test set, in which all gender forms in the reference translation are as in the real data; 2) a gender-reversed test set in which all gender forms in the reference translation are opposite to those in the real data; 3) a test set where all gender forms in the reference translation are masculine; 4) a test set where all gender forms in the reference translation are feminine.

Table 6.6 reports the BLEU scores on all four configurations. Again, a consistent and statistically highly significant ($p < 0.01$) improvement of the gender-aware model over the baseline model is observed. The improvement is particularly high for sentences in the female version with +10.86 BLEU. Overall, the results are similar to those reported in the main results for the EN>RU model (see Section 6.1.1).

The BLEU scores of the baseline model have a maximum difference of 7.68 points (25.72 to 33.40), while the BLEU scores of the gender-specific model are much closer to each other, with a maximum difference of only 1.41 points (36.11 to 37.52). This shows that the gender bias of the baseline model has been significantly reduced in the gender-aware model.

Despite this improvement, the results indicate that a weak gender bias still exists even in the gender-aware model. Accordingly, the BLEU scores of the feminine variants are slightly lower than those of the masculine variants (36.58 vs. 37.17), and likewise, the scores of the gender-reversed variants are lower than those of the original sentences (36.11 vs. 37.52). Ideally, we would expect approximately the same BLEU values for all four test set variants. Note that this result differs from the main results, where the BLEU value was highest for sentences containing feminine forms.

On a side note, the results for the baseline model show that the model's tendency to stereotypically associate certain contexts with certain genders is relatively weak. The BLEU score of the gender-reversed test set variant

test set	# tags	baseline				gender-aware			
		m	f	n/b	error rate	m	f	n/b	error rate
all masculine	205	134	23	48	11.22%	171	1	33	0.49%
all feminine	205	134	23	48	65.37%	2	171	32	0.98%

Table 6.7: Confusion matrix of gender in the reference translations and detected gender in the hypotheses (male, female, none/both) for EN>RU on 200 sentences that differ only in gender and are otherwise equivalent. The error rate is the percentage of opposite-gender forms out of the total number of tags.

is only slightly below the original variant (29.14 vs. 30.14). On the other hand, the baseline model's tendency to favor the masculine forms in all contexts is quite strong, as indicated by the 7.68 BLEU point difference in performance between the feminine and masculine variants (25.72 vs. 33.40).

Table 6.7 reports the results of the evaluation of gender correctness. Again, it can be seen that the error rate is greatly reduced for both genders. For the sentences with female forms, the error rate of the baseline model reduces from 65.37% to 0.98%. Overall, these results confirm those in the main evaluation.

6.2 Random tagging

Table 6.8 reports the BLEU scores for four different configurations of random tagging with the EN>RU model and the EN>FR model: 1) no input tags; 2) balanced input tags, 50% male and 50% female; 3) only masculine tags; 4) only feminine tags. In addition, the table also shows the confusion matrix of gender forms in the hypothesis and the error rate of opposite gender forms.

The differences in BLEU scores for EN>RU between different configurations for all three tagging approaches are not statistically significant ($p > 0.05$), as desired. As for EN>FR, the same is true for all configurations of random tagging among each other, but not compared to the tag-free usage (44.83 on average compared to 46.05 with no input tags).

With the exception of this difference in EN>FR, random tagging works as expected. First, the BLEU scores are similar among each other, despite the fact that the reference contains fixed gender forms. Second, the number of male and female forms in the hypotheses is approximately balanced

model	config	BLEU	m	f	n/b	error rate
	reference	–	784	499	8 738	–
	w/o tags	30.47	657	228	9 115	2.15%
EN>RU	balanced	30.52	848	802	8 377	0.23%
	all masculine	30.58	1 557	32	8 411	0.32%
	all feminine	30.32	46	1 598	8 357	0.46%
	reference	–	501	218	9 284	–
	w/o tags	46.05	314	89	9 598	1.13%
EN>FR	balanced	44.87	299	231	9 471	0.34%
	all masculine	44.88	501	27	9 473	0.27%
	all feminine	44.73	67	457	9 476	0.67%

Table 6.8: BLEU scores and detected gender (male, female, none/both) in the hypotheses for four different configurations of random tagging. The error rate reports the number of opposite-gender forms for single-gender configurations and the number of samples needed to be different to have an equal number of forms per gender for mixed-gender configurations.

with the gender-balanced input tags, showing 848 male forms vs. 802 female forms in the case of EN>RU and 299 male forms vs. 231 female forms in the case of EN>FR. And third, in the single-gender configurations, only a small number of hypotheses contain forms of the other gender.

This is reflected in the error rate. For the single-gender configurations, the error rate is calculated as the proportion of opposite-gender forms in the output. For the tagless and balanced configurations, the error rate is calculated as the number of gender forms in the output that should be in the opposite gender to achieve a fully balanced result. The error rates show that the gender-aware models are successful both in generating consistent gender forms across an entire set of sentences and in generating balanced gender forms, successfully overcoming the gender bias inherent in most machine translation models. These two observations correspond to the two use cases discussed in Chapter 4.

It should be noted that the confusion matrix is not completely reliable, since not all gender mismatches between the input tags and the gender in the output are actually errors of the model. First, some alleged mismatches can be attributed to errors of the rule-based annotation scripts. These mismatches are uninteresting, but it should be kept in mind that they slightly distort the evaluation result. Second, gender mismatches in

model	type	text	gender
EN>RU	input	<1m> *I'm his mother.*	?
	hypothesis	*Я его мать.*	f
EN>FR	input	<1f> *Because I'm a man.*	?
	hypothesis	*Parce que je suis un homme.*	m

Table 6.9: Examples of how input tags are ignored when words are present that already convey a gender that is contradictory to the input tags.

model	type	text	gender
EN>RU	input	<1m> *It's so good to see you!*	m
	hypothesis	*Я так **рада** тебя видеть!*	f
	expected	*Я так **рад** тебя видеть!*	m
EN>FR	input	<1f> *I was impolite.*	f
	hypothesis	*J'ai été **impoli**.*	m
	expected	*J'ai été **impolie**.*	f

Table 6.10: Examples of incorrect gender forms in the model output.

some cases arise from the fact that the input tag is in contradiction with a word in the input sentence that already reveals the gender of the speaker or addressee. Table 6.9 gives two examples of such sentences. These examples show that the model is to some extent able to ignore conflicting input tags if the sentence itself already sufficiently expresses the gender of the speaker or addressee, something that is arguably desirable.

The remaining gender mismatches in the confusion matrix are true errors of the model, where an input tag is not appropriately reflected in the generated translation. Table 6.10 contains two examples of such incorrect translations.

6.3 Tag prediction in triangular translation

6.3.1 Evaluation with BLEU

Table 6.11 reports the results for the Russian-English-Russian (RU>EN>RU) back-translation model on the full test set of 10k samples and on different subsets of the test set. The gender-aware model outperforms the base-

test set	# samples	baseline	w/o tags	w/ tags	improvement
full test set	10 000	43.38	42.29	**45.74**	+2.36
cont. no gender/TV	6 498	43.53	43.65	**43.65**	+0.12
cont. gender/TV	3 502	43.07	40.61	**48.33**	+5.26
cont. gender	1 262	43.14	37.05	**49.43**	+6.29
cont. masculine	784	46.04	38.13	**49.57**	+3.53
cont. feminine	499	38.52	35.11	**49.90**	+11.38
cont. only TV	2 240	43.03	42.95	**47.59**	+4.56
cont. T	1 309	48.64	47.47	**49.40**	+0.76
cont. V	931	35.58	36.98	**45.23**	+9.65

Table 6.11: BLEU scores for RU>EN>RU on various subsets of the test set. Best values in bold. The last column shows the improvement of oracle tagging (w/ tags) over the baseline model.

line model on all subsets, and the improvement is statistically significant ($p < 0.05$) on the full test set and on all subsets containing gender or politeness expressions. With both gender and politeness level, it is again observed that the improvement of the gender-aware model over the baseline is particularly high for the variant which is found less often in the training data and thus disfavored by the translation model. The improvement is as high as +9.65 BLEU for samples containing a V-form, and +11.38 BLEU for samples containing a reference to a female speaker or addressee. This indicates that tag prediction is particularly beneficial for variants disfavored by the model.

The back-translation model consists of two sub-models, the RU>EN model, which also predicts the appropriate tags in the English output, and the EN>RU model, which has already been evaluated in section 6.1.1. The overall gain of +2.36 BLEU points of the gender-aware back-translation model is almost identical to the gain of +2.34 of the EN>RU model itself (see 6.1.1 section). Moreover, the improvements across the different subsets of the test set are similar to those observed for the EN>RU model. This shows that the RU>EN model is able to reliably predict the tags when translating into English.

The BLEU values for the RU>EN>RU model are much higher than those for the EN>RU model itself, although the latter is used here as a sub-model. This may be due to the model's tendency to translate more literally com-

test set	# samples	baseline	w/o tags	w/ tags	improvement
full test set	10 000	44.87	44.37	**47.61**	+2.74
cont. no gender/TV	6 994	46.98	46.65	**46.74**	-0.24
cont. gender/TV	3 006	39.91	38.98	**49.63**	+9.72
cont. gender	658	40.94	37.22	**51.86**	+10.92
cont. masculine	336	43.29	40.61	**49.96**	+6.67
cont. feminine	327	38.25	32.76	**53.86**	+15.61
cont. only TV	2 348	39.58	39.53	**48.91**	+9.33
cont. T	1 314	38.26	37.33	**48.08**	+9.82
cont. V	1 034	41.28	42.10	**49.56**	+8.28

Table 6.12: BLEU scores for RU>EN>FR on various subsets of the test set. Best values in bold. The last column shows the improvement of oracle tagging (w/ tags) over the baseline model.

pared to human reference translations. Compared to the actual English sentences in the EN-RU test set, the English translations generated by the RU>EN model are likely to be closer to the Russian source sentences and can therefore be reconstructed with higher accuracy.

Table 6.12 reports the results for the Russian-English-French (RU>EN>FR) triangular translation model on the full test set of 10k samples and on different subsets of the test set. The gender-aware model outperforms the baseline model on all subsets except for the subset of samples containing no gender or politeness expressions, where a slight decrease of -0.24 BLEU points is observed. This decrease is not statistically significant ($p > 0.05$), but the improvement on all other subsets is ($p < 0.05$).

The triangular translation model consists of two submodels, the RU>EN model, which also predicts the appropriate tags in the English output, and the EN>FR model, which has already been evaluated in section 6.1.1. The overall gain of +2.74 BLEU points of the gender-aware triangular translation model is slightly higher than the gain of +2.14 of the EN>FR model itself (see 6.1.1 section).

The BLEU scores themselves are close to those observed for the EN>FR model, with 47.61 on the full test set for the RU>EN>FR model as compared to 48.10 for the EN>FR model. It should be noted though, that unlike in the back-translation model, the test sets used are not the same. The test set used here is from Tatoeba. In theory, this test set is thus slightly out-

of-domain (being based on Tatoeba instead of OpenSubtitles), however, my personal impression is that the Tatoeba sentences are overall simpler than the ones in the OpenSubtitles corpus. I any case, BLEU scores between the two test sets are not directly comparable. Nevertheless, the improvements on the different subsets of the test set are similar to those observed for the EN>RU model. For samples containing masculine forms, the improvement of tag prediction over the baseline is +6.67 BLEU, and for samples containing feminine forms, it is as high as +15.61 BLEU. These positive results indicate that the tag prediction approach works very well.

6.3.2 Comparison with DeepL

test set	# samples	DeepL	gender-aware	difference
full test set	10 000	**50.60**	47.61	-2.99
cont. no gender/TV	6 994	**53.40**	46.74	-6.67
cont. gender/TV	3 006	45.00	**49.63**	+4.63
cont. gender	658	42.96	**51.86**	+8.90
cont. masculine	336	**50.06**	49.96	-0.10
cont. feminine	327	34.42	**53.86**	+19.44
cont. only TV	2 348	45.61	**48.91**	+3.30
cont. T	1 314	43.84	**48.08**	+4.24
cont. V	1 034	47.87	**49.56**	+1.69

Table 6.13: Comparison of BLEU scores with DeepL for RU>FR on various partitions of the test set.

In gender-aware triangular translation, tags are used exclusively internally within the model and there are no tags in the input. This allows a direct comparison between the gender-aware RU>EN>FR model and the Russian–French model from DeepL. Table 6.13 reports the results for the Russian–French model of DeepL on the RU-FR test set, and for comparison, the results of the RU>EN>FR triangular translation model on the same test set. The BLEU scores of the RU>EN>FR model are identical to those in Table 6.12 and are repeated here for convenience.

The BLEU scores for DeepL on different subsets of the test set reveal that the Russian–French model of DeepL has a problem with gender and politeness forms. DeepL has a BLEU score of 53.40 for samples that do

not contain a gender or politeness expression, while the BLEU score for samples that do contain a gender or politeness expression is considerably lower at only 45.00. Similarly, the scores for samples containing gender expressions reveal that the model of DeepL has a strong gender bias towards the male gender, resulting in a score of 50.06 for samples containing masculine forms as opposed to only 34.42 for samples containing feminine forms.

These issues indicate that the Russian–French model of DeepL was not, or not exclusively, trained on parallel Russian–French data, because in this scenario gender would not be lost in translation. Therefore, DeepL is likely to use either a triangular model or an English-centric multilingual model for this language pair. In any case, the BLEU scores demonstrate that gender and politeness expressions are a real obstacle to the Russian–French model of DeepL.

The results for the RU>EN>FR model show that the tag prediction approach proposed in this paper successfully solves this problem. On the full test set, the RU>EN>FR model is -3.99 BLEU points behind DeepL, and this difference is even more pronounced on samples without a gender or politeness expression, with -6.67 BLEU points. However, on samples containing a gender or politeness expression, the balance is reversed, and the RU>EN>FR model outperforms DeepL by +4.63 BLEU points, and this improvement is statistically significant ($p < 0.05$).

An improvement is obtained both for samples containing a gender expression (+8.90) and for samples containing only a politeness expression (+3.30). Crucially, however, the improvement in samples containing a gender expression is entirely due to better performance in samples containing feminine forms. For samples containing masculine forms, the RU>EN>FR model is actually slightly behind the Russian-French model of DeepL, but the difference of -0.10 points is not statistically significant ($p > 0.05$). For samples containing feminine forms, however, the RU>EN>FR model outperforms DeepL by +19.44 BLEU points, a result which is statistically highly significant ($p < 0.01$) and impressively illustrates how much DeepL neglects feminine forms.

DeepL is generally considered a high-quality machine translation service. In light of this, the results are very encouraging. The comparison with DeepL shows that automatic tag prediction is an effective method to solve the problem of gender bias in triangular translation.

7 Discussion

7.1 General observations

As stated in Chapter 1, there are two key questions regarding the gender tag approach to machine translation: 1) are the models able to generate translations that take into account one or more side constraints, and 2) how does the use of tags affect the overall translation quality in terms of BLEU?

The answer to question 1 is a clear yes. The experimental results show that it is indeed possible to control the desired attributes with high accuracy, as reported in Section 6.1.2. Only 3 of a total of 1298 gender references (0.23%) in the EN-RU test set were detected as having an incorrect gender form. Likewise, only 4 of a total of 724 gender references (0.55%) in the EN-FR test set had an incorrect gender form. In addition, in interactive testing, I did not observe any spurious generalization of tags to entities other than the intended ones. For example, a 2nd person gender tag does not affect the gender of words that refer to the 1st or a 3rd person.

Regarding question 2, the higher accuracy of gender-specific inflection is consistently reflected in higher BLEU scores for the gender-aware model. An improvement over baseline of +2.34 for EN>RU and of +2.14 for EN>FR is observed on the full test sets. As expected, the improvement is higher for the relevant subset of samples exhibiting at least one of the considered phenomena, namely +4.64 for EN>RU and +5.57 for EN>FR. All improvements in BLEU scores are statistically significant.

It is worth noting that some researchers have found that improved gender accuracy does not necessarily improve BLEU scores, and may sometimes even lead to a deterioration (e.g., Rabinovich et al. [2017], Vanmassenhove et al. [2018]). For those reporting an increase in BLEU, the increase is typically smaller than that found here. For example, Elaraby

et al. [2018], whose work is closest to this one, report an improvement over the baseline model of +0.58 BLEU on the full test set and +2.14 BLEU on the gender subset. The test sets are not identical, to be sure, but both works use data from OpenSubtitles.

To give an example of the capacity of gender-aware models, I would like to return to the example presented in Chapter 1. Table 7.1 shows the translation hypotheses generated by the EN>FR model for a single English source sentence with eight different tag configurations:

input		hypothesis
<1m> <2m> <T> *I'm glad you're my friend.*	→	*Je suis content que tu sois mon ami.*
<1f> <2f> <T> *I'm glad you're my friend.*	→	*Je suis contente que tu sois mon amie.*
<1m> <2f> <T> *I'm glad you're my friend.*	→	*Je suis content que tu sois mon amie.*
<1f> <2m> <T> *I'm glad you're my friend.*	→	*Je suis contente que tu sois mon ami.*
<1m> <2m> <V> *I'm glad you're my friend.*	→	*Je suis content que vous soyez mon ami.*
<1f> <2f> <V> *I'm glad you're my friend.*	→	*Je suis contente que vous soyez mon amie.*
<1m> <2f> <V> *I'm glad you're my friend.*	→	*Je suis content que vous soyez mon amie .*
<1f> <2m> <V> *I'm glad you're my friend.*	→	*Je suis contente que vous soyez mon ami.*

Table 7.1: Translations by the EN>FR model for the same English source sentence with eight different tag configurations. All translations are correct.

This example illustrates the fine-grained control that is possible with the gender-aware models. All translation variants are correctly predicted by the model in a controlled manner.[1] But not only that: the model also correctly predicts all variants where only two tags are specified (where the model is free to choose the value for the missing parameter), and likewise all variants where only one tag is specified. For example, *<2f> I'm glad you're my friend* is translated into *Je suis content que vous soyez mon amie*, in which the word *amie* is feminine as expected. When no tags are used with this sentence, the model produces the masculine form for both persons and the formal level of politeness. This corresponds to those forms that are more frequent in the EN-FR training corpus. As a side note, and as mentioned in Chapter 6, the model does not rely on the order of the tags and can also correctly interpret tags in a different order than the one in Table 7.1.

It is interesting to note that the results have been very consistent for both language pairs (EN>RU, EN>FR). This is not self-evident because,

[1] The only difference from the 'reference translations' I gave in Chapter 1 is that the word *content/-e* is used throughout instead of *heureux/-euse*, but both can be considered correct.

first, French and Russian differ in their patterns of using gender morphology, and second, the annotation of gender and politeness tags was done with two independent scripts that are not necessarily of the same quality. Based on these results, it is reasonable to assume that comparable results can be obtained for other languages that exhibit similarly extensive gender morphology, notably the Romance, Slavic, and Semitic languages.

An interesting byproduct of this work is the two heuristic scripts for gender annotation in Russian and French. Originally designed only to annotate the training corpus, they proved useful for also detecting and quantifying gender bias in the data. For example, these scripts revealed that in the English-French and English-Russian corpora from OpenSubtitles, only about one-third of all speakers and addressees are female, while two-thirds are male. Interestingly, Vanmassenhove et al. [2018] found the same distribution in the Europarl corpus, based on their analysis of the metadata. The annotation scripts can be applied to any data set in Russian or French to obtain a quick and rough estimate of the gender distribution in the corpus. In addition, the annotations make it possible to extract a balanced sample of sentence pairs that can be used, for example, as a test or validation set.

7.2 Applications of gender tagging

7.2.1 Reducing post-processing effort

Gender-aware machine translation has several possible applications. First and foremost, it enables the user to actively control the desired attributes in the translation. This may be particularly useful for professional translation service providers, where automatic translation is used in combination with manual post-editing. Gender and politeness are frequently occurring phenomena, and their mistranslation can cause a lot of tedious post-editing work. In fact, as mentioned earlier, a survey by Etchegoyhen et al. [2014] shows that post-editors find it frustrating to repeatedly correct politeness levels in translation, and it is reasonable to assume that this also applies to correcting inappropriate gender forms.

Besides translators and post-editors, non-professional users could also benefit from this feature. Importantly, a person's gender is a concept that everyone is familiar with, and users do not need to be linguistically

trained or familiar with the grammar of the target language to understand the meaning of these attributes. Users also do not need to know which words are affected by their choice, nor do they need to decide whether the tags are needed in the translation or not. The models can handle both superfluous and missing tags, making tagging suitable for real-world applications.

Moreover, in many text genres, be it speeches, letters or blog posts, speaker and addressee usually do not change across sentence boundaries, but remain constant throughout the text. Therefore, in many cases, this allows for very efficient control of gender forms across an entire document with just a single document-level gender specification.

From a user experience point of view, requiring users to type in the tags is not an optimal solution. First, users would have to know the exact form of all possible tags, which is unrealistic. Second, this would introduce the risk of typing errors which can lead to unexpected output. And third, the repeated application of tags via typing can quickly become tedious as well. Instead, the control of the attributes can and should be implemented as a graphical interface, e.g. as a button or a drop-down menu. DeepL already offers such a drop-down menu for some selected languages to allow control over the politeness level. Users can select either the formal or informal level of politeness, or leave it unspecified. Similar buttons can be added for speaker gender and addressee gender. Other solutions are also possible, such as an interactive approach where the user is prompted to specify one or more of the attributes if this information is needed for the translation.

7.2.2 Mitigating gender bias

As discussed in Section 3.1, all major commercial machine translation models were reported to exhibit a gender bias, generating masculine forms much more often than feminine ones in most ambiguous contexts. The reasons for this are both unbalanced training data and an over-generalization of the models towards the more frequently attested forms.

Gender-aware machine translation offers the possibility to mitigate this bias. Assuming a default situation where the user does not or cannot specify a gender, the model can generate different translation variants and display them side by side. Google Translate offers this solution for third-

person gender,[2] but so far it is limited to some language pairs and to single-sentence inputs. Displaying multiple translation variants side by side is an attractive solution, but it also has its limitations. With the three tag sets used in this work, up to eight different and equally correct translation variants can be generated in some cases, and this does not yet include the third-person gender or any other additional attribute. Generating eight different variants is computationally expensive and displaying them in a user-friendly way is a challenge. This is particularly true for longer source texts. However, parallel display is a good default setting for relatively simple inputs that successfully mitigates gender bias.

Another solution to mitigate gender bias with gender-aware models that may be particularly interesting in more complex cases is balanced random tagging (see Section 5.2). In this approach, male and female tags are randomly applied to the user input with equal probability. As shown in Section 6.2, this approach can successfully reduce the gender bias in the output, even if the underlying translation model has a gender bias. This approach ensures that a female variant occurs in about half of the cases and a male variant in the other half. It does not require user action or the expensive generation of multiple translation variants. When used as a default in combination with a menu for manually specifying gender and politeness values, it can be accompanied by a remark indicating the automatically selected gender and offering the possibility to switch if desired.

In some situations, the gender of the speaker or the addressee can also be inferred from the context or from metadata. For example, many social media platforms offer the option to translate posts in foreign languages with a single click. In such cases, if the gender of the person who created the post can be retrieved from that person's profile information, it can be used to create an adequate translation with the correct gender for the 1st person.

7.2.3 Improving text coherence

Standard machine translation models translate each sentence in isolation, which can lead to inconsistencies in word choice throughout an entire text. This also applies to politeness and gender values, which usually should remain constant across different sentences. In sentence-level machine

[2] https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html

translation, the values for gender and politeness in the output are not only unpredictable, but can also vary from sentence to sentence, which can be very confusing for the reader and reduce the perceived quality of the automatic translation.

Gender and politeness tagging provides a simple way to mitigate this problem by applying the same tag configuration to every sentence in a text. If a tag configuration is user-supplied or can be inferred from context, applying it to the entire text rather than a single sentence will improve text coherence in many cases. However, the approach is not limited to user-supplied or context-induced tags, but also works with random tagging. If the gender of the speaker, the gender of the addressee, or the social situation is not known, either gender or politeness level can be used in the translation, but a consistent choice increases text coherence and makes the text more natural and readable.

7.2.4 Improving triangular translation

One particularly promising application of tags is tag prediction in triangular translation. In triangular translation, a source sentence is not translated directly into the target language, but first into a pivot language and in a second step from there into the target language. English is often used as a pivot language in triangular translation, and since it lacks gender and politeness morphology, this information is lost in the intermediate English translation and therefore cannot be successfully transferred to the target language.

In Chapter 5, I proposed to train the source-to-pivot model in such a way that it outputs tags indicating the gender and politeness values present in the source sentence along with the English translation. This tag-enriched intermediate translation is then translated into the target language using a pivot-to-target model that can interpret tags. To my knowledge, this is a novel approach that has not been proposed to date.

As described in Section 6.3, the experimental results for this approach have been very positive. A direct comparison with DeepL for translation from Russian into French has shown that the RU>EN>FR triangular model significantly significantly outperforms DeepL for sentences containing gender and/or politeness expressions. For sentences containing female references, the triangular model is +19.44 BLEU points ahead of DeepL. Not only do the tags improve the BLEU score, however, but they also success-

fully eliminate gender bias. While the DeepL model was found to perform significantly worse on sentences with a female reference, no such difference is observed in the model proposed here.

These results have implications that go beyond gender. They indicate that translation through a pivot language can be greatly improved by using automatically predicted tags that compensate for what is lost in the pivot language. This may include other phenomena, such as the original word order or the distinction between singular and plural 'you'.

7.3 Remaining issues and future work

The tagging approach has proven effective overall, allowing fine-grained control of multiple attributes in translation with high precision. However, its main weakness is that translations of the same source sentence in combination with different tags may differ in words that are not directly related to the controlled attributes. For example, the model might translate the English sentence *I am happy* into *Je suis content* for the male gender and into *Je suis heureuse* for the female gender. Both *content* and *heureuse* can be considered correct translations of *happy*, but the inconsistency of word choice between the male and female versions is probably undesirable.

To get a rough estimate of how common such discrepancies are, I performed an evaluation on the samples from the English-Russian test set that contain at least one relevant gender form. For this, I translated the sentences once with the 1st and 2nd person gender set to male and once with both set to female. Then, for each sentence, I evaluated whether both translation variants have the same number of words and whether each word begins with the same letter. This works because gender inflection in Russian is found almost exclusively at the end of a word.[3] This quick analysis revealed that about 20% of the hypotheses exhibit differences unrelated to gender, i.e., differences in word choice or syntax.

Opinions differ on whether such discrepancies are desirable or not. For example, Rabinovich et al. [2017] actively advocate for the preservation of features of gender-specific language use in translation. I would argue

[3] A notable exception are the words *∂pyr* (*male friend*) and *no∂pyra* (*female friend*). Also, this simple heuristic fails to detect different lemmas that coincidentally start with the same letter.

instead that different translations by gender are undesirable because the association of language style with gender reproduces gender stereotypes.

However, it is not trivial to achieve equivalence across different translation outputs with the tagging method used in this work. A possible solution to this problem is to use a standard machine translation model instead and to perform the gender adaptation only in a post-processing step. This two-step approach is less elegant, but probably more reliable for ensuring consistency among translation variants. Indeed, Google has switched to this approach recently (see the footnote in Section 7.2.2). In any case, a systematic comparison between the post-processing approach and the end-to-end approach used in this work is definitely worth further investigation.

Another interesting direction for further research is the inclusion of additional constraints. To begin with, side constraints may be extended to the gender of third-person references. However, many other constraints are conceivable, e.g., to control tense, aspect, voice, the length of the output sequence, or a particular terminology to be used in the translation. Most of these side constraints have been implemented in previous research, but usually only in isolation. There has not yet been much research on how reliably machine translation models can deal with a larger number of constraints at the same time. In this respect, it is certainly worthwhile to further investigate the vector intervention approach proposed in Schioppa et al. [2021] as an alternative to tagging.

In another direction, the promising results obtained in this work for tag prediction in triangular translation suggest that this method is worth further investigation. As triangular models are increasingly replaced by multilingual models, further research could explore how tags could be used in multilingual models to improve translation quality between zero-shot or few-shot language pairs.

8 Conclusion

The aim of the present work has been to develop gender-aware machine translation models for English–Russian and English–French that are able to take into account side constraints related to the gender of the speaker, the gender of the addressee, or the desired level of politeness. The main contributions of this work are the creation of two gender-annotated corpora for English-Russian and English-French, the training of models on the annotated data, and experiments to evaluate the performance of gender and politeness tagging.

The experimental results were highly positive. Oracle experiments that simulate the real-world use case of a user who desires to have a particular gender in the translation show highly significant improvements in terms of BLEU compared to the baseline model. Furthermore, the evaluation of gender accuracy showed that the models successfully translate side constraints into the corresponding target-side morphology. This capability is not limited to single constraints, but also works for up to three simultaneous constraints. In addition, the models can handle both superfluous and missing tags.

A particularly high improvement was obtained in both language pairs for samples containing a reference to a female person. Because feminine forms are underrepresented in the training data, most machine translation models, including commercial ones, produce masculine forms much more often than feminine forms in ambiguous contexts. However, gender bias in machine translation can be successfully mitigated with the tagging approach, both by allowing the user to control the resulting gender and by using random gender tags with equal probability if the user does not wish to provide this information. Unlike all major commercial machine translation systems, the gender-aware models used in combination with balanced random tagging produce approximately equal proportions of female and

male forms in the output. In addition, the consistent use of tags throughout a text can also be beneficial for text coherence, as it suppresses random gender variations across sentences, which can be very confusing for the reader.

A particularly interesting finding of this work is that gender and politeness tags can also be automatically predicted in triangular translation to counteract the loss of information in the pivot language. Given the high accuracy that neural machine translation has achieved today, the main reason for the performance decrease in triangular translation is arguably not error propagation, but mainly information loss. This loss of information can be prevented by allowing the model to learn to append the information that would otherwise be lost in the pivot language. This method is not limited to gender and politeness, but can be extended to any other relevant feature.

In summary, gender-aware machine translation has many potential applications. In particular, it may be useful for:

- reducing tedious and repetitive manual post-processing work
- mitigating gender bias with balanced random tagging
- increasing text coherence with consistent tagging across sentences
- improving translation quality in triangular translation via automatic tag prediction

Possible future research directions include the implementation of a suitable solution for gender references in the 3rd person as well as further exploration of multi-constraint machine translation. In addition, based on the good results, tag prediction as a method to minimize information loss in triangular translation deserves further attention. Furthermore, the applicability of this method to multilingual models is worth investigating.

List of Tables

References

J. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *Advances in NIPS 2016 Deep Learning Symposium*, 2016.

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

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, 2020.

T. Bergmanis, A. Stafanovivcs, and M. Pinnis. Mitigating gender bias in machine translation with target gender annotations. In *WMT*, 2020.

I. Caswell, C. Chelba, and D. Grangier. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy, 2019.

W. I. Cho, J. W. Kim, S. M. Kim, and N. S. Kim. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy, 2019.

P. K. Choubey, A. Currey, P. Mathur, and G. Dinu. GFST: Gender-filtered self-training for more accurate gender in translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1640–1654, Punta Cana, Dominican Republic, 2021.

M. Costa-jussà, C. Escolano, C. Basta, J. Ferrando, R. Batlle, and K. Kharitonova. Gender bias in multilingual neural machine translation: The architecture matters. *Arxiv Preprint*, 2020.

M. Costa-jussà, C. Basta, and G. I. Gállego. Evaluating gender bias in speech translation. *Arxiv Preprint*, 2021.

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, 2019.

M. Elaraby, A. Y. Tawfik, M. Khaled, H. Hassan, and A. Osama. Gender aware spoken language translation applied to English-Arabic. In *2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, 2018.

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

T. Etchegoyhen, L. Bywood, M. Fishel, P. Georgakopoulou, J. Jiang, G. van Loenhout, A. del Pozo, M. S. Maučec, A. Turner, and M. Volk. Machine translation for subtitling: A large-scale evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 46–53, Reykjavik, Iceland, 2014.

A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research 22*, pages 1–48, 2021.

W. Feely, E. Hasler, and A. de Gispert. Controlling Japanese honorifics in English-to-Japanese neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 45–53, Hong Kong, China, 2019.

M. Gaido, B. Savoldi, L. Bentivogli, M. Negri, and M. Turchi. Breeding gender-aware direct speech translation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3951–3964, Barcelona, Spain, 2020.

N. Habash, H. Bouamor, and C. Chung. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy, 2019.

M. Honnibal and I. Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

C. Kobus, J. Crego, and J. Senellart. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 372–378, Varna, Bulgaria, 2017.

T. Kocmi, T. Limisiewicz, and G. Stanovsky. Gender coreference and bias evaluation at WMT 2020. In *WMT*, pages 357–364, 2020.

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, 2005.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, 2007.

J. Kuczmarski and M. Johnson. Gender-aware natural language translation. In *Technical Disclosure Commons*, 2018.

T. Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, 2018.

T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, 2018.

S. M. Lakew, M. A. D. Gangi, and M. Federico. Controlling the output length of neural machine translation. *Arxiv Preprint*, 2019.

P. Lison and J. Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 923–929, Portorož, Slovenia, 2016.

P. Lison, J. Tiedemann, and M. Kouylekov. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, pages 1742–1748, Miyazaki, Japan, 2018.

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

M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. Fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, pages 48–53, Minneapolis, Minnesota, 2019.

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, 2002.

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, 2018.

M. O. R. Prates, P. H. C. Avelar, and L. Lamb. Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, 2019.

E. Rabinovich, R. N. Patel, S. Mirkin, L. Specia, and S. Wintner. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain, 2017.

A. A. Rescigno, J. Monti, A. Way, and E. Vanmassenhove. A case study of natural gender phenomena in translation: A comparison of Google Translate, bing Microsoft translator and DeepL for English to Italian, French and Spanish. In *Workshop on the Impact of Machine Translation*, pages 62–90, 2020.

N. Roberts, D. Liang, G. Neubig, and Z. C. Lipton. Decoding and diversity in machine translation. *CoRR*, abs/2011.13477, 2020.

V. M. Sánchez-Cartagena, M. Bañón, S. Ortiz-Rojas, and G. Ramírez-Sánchez. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 955–962, Belgium, Brussels, 2018.

D. Saunders and B. Byrne. Reducing gender bias in neural machine translation as a domain adaptation problem. In *ACL*, pages 7724–7736, 2020.

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

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.

A. Schioppa, D. Vilar, A. Sokolov, and K. Filippova. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Punta Cana, Dominican Republic, 2021.

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*, pages 35–40, San Diego, California, 2016a.

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, 2016b.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

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, 2019.

E. Stergiadis, S. Kumar, F. Kovalev, and P. Levin. Multi-domain adaptation in neural machine translation through multidimensional tagging. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 396–420, 2021.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

S. Takeno, M. Nagata, and K. Yamamoto. Controlling target features in neural machine translation via prefix constraints. In *Proceedings of the 4th Workshop on Asian Translation (WAT)*, pages 55–63, Taipei, Taiwan, 2017.

J. Vamvas and R. Sennrich. Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265, Punta Cana, Dominican Republic, 2021.

E. Vanmassenhove, C. Hardmeier, and A. Way. Getting gender right in neural machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3003–3008, Brussels, Belgium, 2018.

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, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269, 1967.

E. Voita, P. Serdyukov, R. Sennrich, and I. Titov. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia, 2018.

E. Voita, R. Sennrich, and I. Titov. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, 2019a.

E. Voita, R. Sennrich, and I. Titov. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, pages 877–886, Hong Kong, China, 2019b.

H. Yamagishi, S. Kanouchi, T. Sato, and M. Komachi. Controlling the voice of a sentence in Japanese-to-English neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT)*, pages 203–210, Osaka, Japan, 2016.

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, 2018.

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