A taxonomy of bias-causing ambiguities in machine translation

Michal Měchura
NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic
and Fiontar & Scoil na Gaeilge, Dublin City University, Ireland
michmech@lexiconista.com

Abstract

This paper introduces a taxonomy of phenomena which cause bias in machine translation, covering gender bias (people being male and/or female), number bias (singular *you* versus plural *you*) and formality bias (informal *you* versus formal *you*). Our taxonomy is a formalism for describing situations in machine translation when the source text leaves some of these properties unspecified (eg. does not say whether *doctor* is male or female) but the target language requires the property to be specified (eg. because it does not have a gender-neutral word for *doctor*). The formalism described here is used internally by Fairslator[1], a web-based tool for detecting and correcting bias in the output of any machine translator.

1 Introduction: phenomena described by the taxonomy

The taxonomy we are going to introduce in this paper is based on the assumption that biased translations are *always* the result of unresolvable ambiguities in the source text. We will start by demonstrating on a few examples what exactly we mean by ambiguity, what makes ambiguities resolvable or unresolvable, and how the unresolvable ones inevitably lead to biased translations. This will serve as an informal introduction before we proceed to a more formal specification of everything in the rest of the paper.

When translating a sentence such as *she is a doctor* from English into a language such as German which has no gender-neutral word for *doctor*, the translator (machine or human) can translate *doctor* either as male *Arzt* or as female *Ärztin*. The word *doctor* is **ambiguous** for the purposes of this translation. However, the presence of the female pronoun *she* should be enough to tip any well-trained machine translator towards the female reading and to translate *doctor* as *Ärztin* – as indeed most of the major machine translators such as Google Translate and DeepL do. Here, the ambiguity is **resolvable** from context, where by context we mean the rest of the text available to the translator.

Now consider a similar sentence: *I am a doctor*. The word *doctor* is as ambiguous as before, but this time the ambiguity is **unresolvable** from context, as there is no indication anywhere in the text whether the intended referrent of *I* and *doctor* is a man or a woman. In such a situation, the machine translator will typically decide for the male translation because that is what has been seen most often in similar contexts in its training data. This is another way of saying that the machine is making an **unjustified assumption**: unjustified because unsupported by anything actually present in the text being translated. There are two possible ways to "read" the ambiguous word *doctor*, but translations produced by this machine will be consistently biased in favour of the male reading whenever context allows both readings.

Unresolvable ambiguities do not simply happen arbitrarily and unexpectedly. Many kinds of unresolvable ambiguities tend to happen regularly and predictably when certain words occur in the source text inside certain lexicogrammatical patterns, for example *I am a...* or *you are a...* followed by a gender-neutral noun known to have two gender-specific translations in the target language. Fairslator is a tool which detects such patterns and acts on them: it asks the human user to disambiguate (eg. to tell us whether they want the male or female reading) and then it re-inflects the translation accordingly. To enable all this functionality, Fairslator has inside itself a taxonomy for describing *how* the source text is ambiguous and *which way* the human user wants the ambiguity to be resolved. The taxonomy describes the following kinds of unresolvable ambiguities:

- Unresolvable ambiguities in the **gender** of human beings being referred to in the text.

[1] https://www.fairslator.com/

Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 168 - 173
July 15, 2022 ©2022 Association for Computational Linguistics

This covers the well-known case of "occupation words" such as *doctor*, *teacher*, *cleaner*, as well as some less well-known cases such as predicatively positioned adjectives in Romance languages (eg. English *I am happy* → French *je suis heureux* male, *je suis heureuse* female) and verbal participles in Slavic languages (eg. English *I wanted it* → Czech *já jsem to chtěl* male, *já jsem to chtěla* female).

- Unresolvable ambiguities in the **number** of people referred to by the English second-person pronoun *you* (and its possessive companion *your*). For example, in the sentence *you are here* the pronoun *you* has an unresolvable ambiguity (from the perspective of translating it into a language which has separate pronouns for singular and plural *you*) because there is no indication in the text whether the *you* refers to one person or many. (Contrast this with a sentence such as *you are all here* where the ambiguity is resolvable from the presence of the plural word *all*.)

- Unresolvable ambiguities in the **formality** with which people are being addressed in the text. Many European languages have separate second-person pronouns depending on whether the speaker is addressing the listener formally and politely, or informally and casually, eg. French *vous* versus *tu*, German *Sie* versus *du*. An English sentence such as *where are you?* has an unresolvable ambiguity (from the perspective of the target language) because there is no indication in it as to which level of formality is intended, or which level of formality *would* be required if one were speaking in the target language. (Contrast this with a sentence such as *where are you Sir?* where the ambiguity is resolvable from the presence of the formal form of address *Sir*.[2])

As is obvious, the Fairslator taxonomy covers many kinds of translation bias, not just bias in gender, even though gender bias is currently the most vigurously debated kind of bias in machine translation (see Savoldi et al. 2021 for a state-of-the-art survey). In terms of the language categories defined by Savoldi et al. 2021, 3.1.1, the taxonomy can (be adapted to) describe gender bias-causing ambiguities during translation from all *genderless languages* into all *notional gender languages* (languages that encode the gender of humans in pronouns and nouns that refer to them) and *grammatical gender languages* (languages that encode the gender of humans through inflection on words that do not directly refer to humans, such as verbs and adjectives).

That said, the taxonomy in its current incarnation, as presented in this paper, is oriented towards translation from English into other, mainly European, languages, and there is a version of the taxonomy for each **directed language pair**: one for English-to-German, one for English-to-Czech and so on.

2 Bias statement: what is bias in machine translation?

We can now proceed to a more formal definition of what we mean by bias. When we consider machine translation as a black box and simply take its input and output as a pair of texts (the source text in the source language plus the translation in the target language), then we can define the following concepts:

Unresolvable ambiguity A portion of the source text contains an unresolvable ambiguity if, in order to translate it successfully into the target language, some *semantic property* of it needs to be known (such as its gender or grammatical number or level of formality) but this property is *not expressed* in the source text and cannot be inferred from anything in the source text.

Unjustified assumption An unjustified assumption is what happens when, in the face of an unresolvable ambiguity, the machine translator decides for one particular reading of the ambiguous expression over others. The assumption is unjustified because nothing actually present in the source text justifies it. The machine's decision is either random or, if the translator has been constructed through machine learning, predetermined by which reading has been observed more often in the training data.

[2] In fact, the addition of "tag-ons" such as *Sir* or *dude*, such as *he said* or *she said* to the end of the sentence is one method which has been experimented with to "solve" machine translation bias. Effectively, it "tricks" the translator into interpreting things in a particular way. See Moryossef et al. 2019.

Bias A machine translator is biased if, while dealing with unresolvable ambiguities and deciding which unjustified assumptions to make, its decisions are *not* random: it makes certain unjustified assumptions more often than others. For example, if a translator consistently decides for male readings of *doctor* or for singular informal readings of *you* (when these are unresolvably ambiguous in the source text), then the translator is biased.

In other words, we define bias as a purely technical concept, as the tendency of an automated system to make certain unjustified assumptions more often than others. This differs from the popular commonsense understanding of the word *bias* which, in addition to the purely technical process, implies harmful and unjust consequences. This implication is not a necessary part of our definition. Our definition of bias covers bias regardless of whether it is harmful to society (eg. because it perpetuates a stereotype by speaking about doctors as if they must always be men), harmful to an individual (eg. because it offends somebody by addressing them with an inappropriately informal pronoun) or relatively harmless and merely factually incorrect (eg. because it addresses a group of people with a singular pronoun).

Interestingly, our definition applies not only to machines but also to humans: it is not unheard of for human translators to make the same kind of unjustified assumptions and to go about it with the same amount of bias as machines. Good human translators avoid bias by observing the extralinguistic reality (simply *looking* to see eg. whether the speaker seems male or female) and by asking follow-up questions ("what do you mean by *you*?"). Machine translators do not normally have the means to do such things but Fairslator is a plugin which adds the latter ability to any machine translator: the ability to recognize unresolvable ambiguities, to ask follow-up questions, and to re-inflect the translation in accordance with the answers, in a fashion similar to Habash et al. 2019 and Alhafni et al. 2020.

3 Components of the taxonomy

3.1 Axes of ambiguity

To describe the unresolvable ambiguities in a pair of texts (source + translation) in the Fairslator taxonomy, we need to analyze the text pair along three axes:

The speaker axis Is the speaker mentioned in the translation, for example by first-person pronouns? And if so, is the speaker mentioned in the translation in a way that encodes gender, while the source text does not?

The listener axis Is the listener mentioned in the translation, for example by second-person pronouns or implicitly through verbs in the imperative? And if so, is the listener mentioned in the translation in a way that encodes gender, number or formality while the source text does not?

The bystander axis Are any bystanders mentioned in the translation, that is to say, are any people other than the speaker and the listener being referred to by nouns or by third-person pronouns? And if so, are the bystanders mentioned in the translation in a way that encodes gender, while the source text does not?

Each text pair contains zero or one speaker axis, zero or one listener axis, and zero, one or more bystander axes. For each axis, we can use the taxonomy to express the fact that there are or are not any unresolvable ambiguities on this axis, what the **allowed readings** are (eg. the translation can be either masculine or feminine along this axis) and which reading is actually expressed in the translation (eg. the translation is masculine along this axis).

We can illustrate this on an example. Assume the following English sentence and its Czech translation.[3]

> *I would like to ask whether this is your new doctor.*
> *Chtěla bych se zeptat, jestli tohle je tvůj nový lékař.*

Using the three kinds of axes, we can analyze this text pair as follows.

1. The speaker axis is present here. The speaker is mentioned in the translation with the words *chtěla bych* 'I would like to' where the word *chtěla* is a verbal participle and encodes the speaker as female in gender, while the source text is ambiguous as to the speaker's gender.

[3] The example is a little convoluted. This is necessary in order to demonstrate all three axes.

2. The listener axis is also present here. The listener is mentioned in the translation with the word *tvůj* 'your'. This word encodes the listener as singular in number and addressed informally, while the source text is ambiguous on these things. Neither the source text nor the translation say anything about the gender of the listener.

3. Finally, one bystander axis is present here. The bystander is mentioned in the source text by the word *doctor* and in the translation by the word *lékař*. The word in the translation encodes the bystander as male in gender, while in the source text it is ambiguous in gender.

For each axis, we have stated two things. First, which readings are allowed by the source text, for example "the speaker can be interpreted as male or female". Second, which reading is actually expressed in the translation, for example "the speaker has been interpreted as female".

3.2 Ambiguity descriptors

To describe the possible readings on each axis, the taxonomy uses combinations of one-letter abbreviations such as `m` or `f` for masculine or feminine gender, `s` or `p` for singular and plural number, and `t` or `v` for informal or formal form of address (from the Latin pronouns *tu* and *vos*, as is comon in linguistic literature on this topic). Using this code we can re-express the observations from above more succinctly:

```
1. sm|sf : sf
2. st|sv|p : st
3. doctor : sm|sf : sm
```

Human-readably, this means:

1. The speaker axis can be `sm` (singular masculine) or `sf` (singular feminine). Currently it is `sf` (singular feminine).

2. The listener axis can be `st` (singular informal) or `sv` (singular formal) or `p` (plural).[4] Currently it is `st` (singular informal).

3. The bystander axis identified through the nickname `doctor` can be `sm` (singular masculine) or `sf` (singular feminine). Currently it is `sm` (singular masculine).

Each line is a **descriptor** which describes the unresolvable ambiguity on a given axis. Each descriptor consists of:

- A number to indicate which axis is being talked about: `1` for the speaker axis, `2` for the listener axis, `3` for the bystander axis. Each description can contain zero or one descriptor for the speaker axis, zero or one descriptor for the listener axis, and zero or one or more descriptors for the bystander axis.

- For the bystander axis only: a nickname to identify this bystander axis from other bystander axes in this description. This is usually a word taken from the source text. If there is more than one bystander axis in the text pair (which is rare but happens in sentences such as *the doctor asked the nurse to...*) than they must have different nicknames (eg. `doctor` and `nurse`).

- Codes for all the readings allowed by the source text in this axis, separated by vertical lines, for example `st|sv|p`.

- A code for the reading actually expressed in the translation for this axis, for example `st`.

Fairslator uses a slightly different catalogue of descriptors for each directed language pair. As an example, Fairslator's complete inventory of descriptors for English-to-German is given in the Appendix.

4 How Fairslator uses the taxonomy

The main purpose of the taxonomy is to make it possible for Fairslator to formulate *human-friendly disambiguation questions* for users.[5] Here are some examples of descriptors and the disambiguation questions generated from them.

```
1. sm|sf : sf
```

- Who is saying it?
  - a man
  - a woman (selected)

```
1. pm|pf : pm
```

- Who is saying it?

[4] In the plural, Czech has no distinction between formal and informal registers.

[5] Fairslator's target audience is users who speak the source language but do not speak the target language, or do not speak it well enough to be able to detect and correct biased translations on their own.

  - a group containing at least one man (selected)
  - a group of women

2. `st|sv|p : st`
   - Who are you saying it to?
     - one person
       - addressed informally (selected)
       - addressed formally
     - several people

3. `doctor : sm|sf : sm`
   - Who is the person identified as "doctor"?
     - a man (selected)
     - a woman

Once the human user has made a selection from these options, Fairslator re-inflects the translation in accordance with the user's wishes: changes pronouns and nouns accordingly, changes verbs and adjectives so as not break grammatical agreement, and so on. The details of this process, as well as details of how Fairslator detects unresolvable ambiguities in the first place, are not the subject of this paper but some information about this can be found in Měchura 2022.

5 Discussion: where the taxonomy could be improved

I* versus *we The taxonomy assumes that there is always no more than one speaker axis in each text and that its grammatical number never changes: it is always either *I* or *we* but never both. This means that it cannot handle texts where the speaker refers not only to himself or herself (*I*) but also to a group he or she belongs to (*we*), such as *I think we should...*

Multiple voices While the taxonomy is able to handle texts consisting of multiple sentences without problems, it can only do so on the assumption that the axes remain unchanged throughout the text. When the axes do change, as they do in a dialogue (*How are you? Very well, and you?*), then the taxonomy is currently unable to keep track of "who is who" and wrongly assumes that, for example, the people referred to by *you* are the same person throughout.

Word-sense ambiguities The taxonomy is designed to handle unresolvable ambiguities in three semantic properties: gender, number and formality. In principle, however, *any* semantic property can be affected by an unresolvable ambiguity during translation. So, ideally, word-sense ambiguities of *any* kind should be covered by the taxonomy. One example for many is *river* → French *fleuve* 'large river flowing into the sea' versus *rivière* 'small river flowing into another river'. In a sentence such as *we went for a walk along the river* being translated into French, the sense of *river* is unresolvably ambiguous and, if not disambiguated manually by a human user, the machine's translation is bound to be biased in favour of one sense or the other. See Lee et al. 2016 for an inspiring attempt to remove word-sense bias from machine translation through human-driven word-sense disambiguation.

Gender-neutral language In languages where words come in gendered pairs, such as *teacher* → German *Lehrer* male or *Lehrerin* female, it is sometimes possible to construct a gender-neutral neologism by merging them together, such as *Lehrer:in*, in case a gender-neutral word is required. The same can sometimes be done with pronouns, adjectives, verbal participles and other gendered pairs of words. While such neoforms are pragmatically strongly marked and not all writers and readers like them, they do exist and should therefore be included in the taxonomy as one of not two but three gender values: male, female and gender-neutral.

6 Summary

Machine translation technology is getting better all the time at resolving ambiguities from clues in the context. But some ambiguities can never be resolved in this way because there *are* no clues in the context. To avoid bias during the translation of texts that contain unresolvable ambiguities, we need to build tools which are able to (1) recognize that an unresolvable ambiguity has occured and (2) ask the human user to disambiguate manually.

To be able to build such tools at all, what we need first of all is an expressive formalism for *describing* unresolvable ambiguities. This paper has shown how to construct such a formalism for any directed language pair by analysing the source text and its translation from the point of view of three axes (speaker, listener and bystander) and by describing any unresolvable ambiguities that occur in those axes through descriptors which tell us (1) which readings are allowed by the source text and (2) which one of those readings is actually expressed in the translation.

References

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

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic Gender Identification and Reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Hyoung-Gyu Lee, Jun-Seok Kim, Joong-Hwi Shin, Jaesong Lee, Ying-Xiu Quan, and Young-Seob Jeong. 2016. papago: A machine translation service with word sense disambiguation and currency conversion. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 185–188, Osaka, Japan. The COLING 2016 Organizing Committee.

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. 2019. Filling Gender & Number Gaps in Neural Machine Translation with Black-box Context Injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54, Florence, Italy. Association for Computational Linguistics.

Michal Měchura. 2022. We need to talk about bias in machine translation: the Fairslator whitepaper. Technical report.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Appendix: inventory of descriptors

Here we are going to lay out Fairslator's complete inventory of descriptors for **English-to-German**. Each descriptor describes one type of unresolvable ambiguity which is capable of occurring during translation between these two languages, in this direction. We accompany each descriptor with an example to illustrate the ambiguity.

Speaker axis

`1. sm|sf`
I am the new director.
`sm` *Ich bin der neue Direktor.*
`sf` *Ich bin die neue Direktorin.*

`1. pm|pf`
We are teachers.
`pm` *Wir sind Lehrer.*
`pf` *Wir sind Lehrerinnen.*

Listener axis

`2. ts|vs|tp|vp`
Are these your children?
`ts` *Sind das deine Kinder?*
`vs` *Sind das Ihre Kinder?*
`tp` *Sind das eure Kinder?*
`vp` *Sind das Ihre Kinder?*

`2. ts|vs`
Did you do it yourself?
`ts` *Hast du es selbst gemacht?*
`vs` *Haben Sie es selbst gemacht?*

`2. tp|vp`
Did you do it yourselves?
`ps` *Habt ihr es selbst gemacht?*
`vp` *Haben Sie es selbst gemacht?*

`2. tsm|tsf|vsm|vsf`
Are you the new director?
`tsm` *Bist du der neue Direktor?*
`tsf` *Bist du die neue Direktorin?*
`vsm` *Sind Sie der neue Direktor?*
`vsf` *Sind Sie die neue Direktorin?*

`2. tpm|tpf|vpm|vpf`
Are you teachers?
`tpm` *Seid ihr Lehrer?*
`tpf` *Seid ihr Lehrerinnen?*
`vpm` *Sind Sie Lehrer?*
`vpf` *Sind Sie Lehrerinnen?*

Bystander axis

`3. director : sm|sf`
This is the new director.
`sm` *Das ist der neue Direktor.*
`sf` *Das ist die neue Direktorin.*

`3. teachers : pm|pf`
These are our teachers.
`pm` *Das sind unsere Lehrer.*
`pf` *Das sind unsere Lehrerinnen.*