Context-aware and gender-neutral Translation Memories

[1]**Marjolene Paulo,** [1]**Vera Cabarrão,** [2,3]**Helena Moniz**
[2,3]**Miguel Menezes,** [1]**Rachel Grewcock,** [1]**Eduardo Farah**
[1] Unbabel, Lisbon, Portugal
[2] University of Lisbon, Portugal
[3] INESC-ID, Lisbon Portugal
{marjolene.paulo.int, vera.cabarrao, helena,
luis.menezes.ext, rachel.grewcock, eduardo.farah}@unbabel.com

Abstract

This work proposes an approach to use Part-of-Speech (POS) information to automatically detect context-dependent Translation Units (TUs) from a Translation Memory database pertaining to the customer support domain. In line with our goal to minimize context-dependency in TUs, we show how this mechanism can be deployed to create new gender-neutral and context-independent TUs. Our experiments, conducted across Portuguese (PT), Brazilian Portuguese (PT-BR), Spanish (ES), and Spanish-Latam (ES-LATAM), show that the occurrence of certain POS with specific words is accurate in identifying context dependency. In a cross-client analysis, we found that 10% of the most frequent 13,200 TUs were context-dependent, with gender determining context-dependency in 98% of all confirmed cases. We used these findings to suggest gender-neutral equivalents for the most frequent TUs with gender constraints. Our approach is in use in the Unbabel translation pipeline, and can be integrated into any other NMT pipeline.

1 Introduction

Translation Memory (TM) servers are dynamic databases that store frequent and high-quality translated segments (Bowker and Fisher, 2010). In the context of Computer-aided translation (CAT) tools and in the Machine Translation (MT) industry, these segments are used as complementary

modules, to optimize the translation process, reducing costs and increasing the speed of translations. A downside of these systems is that they are usually based on sentence-like units. Combining the sentences risks coherence problems at the document level. This is problematic as a high-quality text is a coherent and cohesive unit with ideas presented in a logical way.

Recent work in this field aims to achieve a context-aware MT system by incorporating more context than the current sentence and tackling discourse phenomena across the document.

Our use case is customer-support content: a content type that differs significantly from other types of text since it represents an interaction between an agent and a customer, usually over email or chat messages. This implies that the text contains a lot of first and second person, and pronouns or anaphora. Even third person references can also be gendered in the target languages in question (*e.g.*, "Está encantada?" "Está encantado?" / Are you thrilled?). Since the translation process is carried out sentence by sentence, gender information associated with that of the addressee is easily lost, resulting in agreement problems throughout the document.

We hypothesize that, by using POS information, we can automatically identify context-dependent and context-independent segments. With the results of context identification, we can then transform segments with gender constraints into gender-neutral TUs, so that these can be reused without damaging translation quality through agreement errors.

As such, part of our work is concerned with creating gender-neutral content. Gender inclusivity is a topic of increasing concern in Natural Language Processing (NLP) and Responsible AI, and exist-

ing approaches to tackle it have used neutral pronouns (Sun et al., 2021), or words or expressions that do not require gender-marking at all (Piergentili et al., 2023).

Aside from creating a more inclusive tone with customers, using gender-neutral content also allows us to reuse the same segments without limits, and without compromising meaning or quality. Our goal is to show how this approach, applied here in the customer-support domain, is equally applicable to other industries and to existing NLP pipelines. Similarly, our strategies to identify context patterns can be integrated into any Neural Machine Translation (NMT) pipeline.

2 Related Work

Context is pivotal to understanding a text. Words are bound to fixed semantics, but can acquire distinct meanings in different contexts. One of the major issues faced in translation, language ambiguity (Koehn, 2020), is linked to the innate flexibility of words and sentence structures in terms of their semantic value. Context plays a fundamental role in allowing us to decode ambiguous segments.

The relationship between context and linguistic expression is complex; context influences and is influenced by linguistic expression. As House (2006) describes, the dependency between these two dimensions (form and the *background*) is omnipresent, and is decisive for the construction and recovery of meaning. In our line of investigation, we use this perspective to observe the relationship between context and the linguistic structures within a document. We consider context crucial for an accurate word/sentence interpretation that would otherwise be lost or misunderstood, resulting in ambiguities and nonsensical text.

A text is a linguistic object with properties organised around discursive concepts like textual coherence and cohesion. These allow us to perceive a sequence of sentences as a unity (Bublitz, 2011). However, contemporary MT still very much relies on a sentence-based translation approach, where sentences are translated in isolation, disregarding context and referential dependencies within a document (Bawden, 2018; Gehring et al., 2017; Hieber et al., 2017). Such tradition can lead to inappropriate and erroneous translations, introducing ambiguities, register and gender issues, *inter alia*.

In recent years, the shift towards *document-level translation* or *context-aware machine translation* has sought to overcome the limitation of current state-of-the-art MT models, and to solve intersentential dependencies by taking into account discursive phenomena (Lopes et al., 2020; Yin et al., 2021a).

Demonstrating the importance of context, Guillou et al. (2018) analyzed and evaluated the performance of 16 NMT systems on the translation of pronouns from English to German with a test set with 200 pronouns. The authors found that all of the NMT systems analyzed had a better performance translating pronouns with intersentential reference. In contrast, the translation of anaphoric pronouns, whose reference was within the sentence, was more difficult.

Bawden et al. (2017) created English to French test sets that tackled coreference, lexical coherence and cohesion as context-aware categories, in order to test the performance of a NMT system with a multi-encoder architecture. The results of their experiment showed positive outcomes for both coherence and cohesion, and conversely, less favorable results for coreference.

In Castilho et al. (2021), the author annotated an English-Brazilian Portuguese corpus made of 60 documents with a total of 3,680 sentences, from six different domains: literary, subtitles, news, reviews, medical and legislation. They consider gender agreement, number agreement, lexical ambiguity, reference, ellipsis and terminology as context-aware issues. The aim of the study is to use the corpus as a test set for the evaluation of MT and quality estimation and to perform linguistic analysis of context issues.

In turn, Yin et al. (2021b), developed SCAT (Supporting Context for Ambiguous Translations), a set of English/French bitexts for contextual support, that access the context used for disambiguation through the identification of the position and characteristics of elements composing a referential chain, thus allowing for the creation of future models that use context effectively. To create the test set, the authors requested annotations from professional translators on the context they considered relevant to resolve intersentential ambiguities in translation.

To generalize, we can state that most relevant work in the domain focuses on the same discursive parameters, such as i) coreference and anaphora resolution (tracing back the referents of

previously-mentioned entities); ii) lexical cohesion (investigating how the different cohesion devices that occur beyond the sentence level are correct); and iii) discourse connectives (exploring how the translation of these words affects the interpretation of the text (Cai and Xiong, 2020; Müller et al., 2018; Yin et al., 2021a; Jwalapuram et al., 2020; Voita et al., 2018)). Regarding our approach to use POS for gender-neutral and for minimizing context-dependent TUs, to the best of our knowledge, this is the first attempt of its kind.

3 Methodology

As previously mentioned, the main goal of context-independent translations is that they can be reused without compromising the coherence of a whole document. In our approach to automatically detect context, we started by classifying a dataset of very frequent TUs from various clients as context-dependent or context-independent, assessing full documents. This dataset comprised 6,368 TUs for PT; 28,604 TUs for ES; 33,623 TUs for PT-BR and 10,026 for ES-LATAM. Due to the high volume of entries, we selected only a sample of the original dataset for analysis: 1,300 entries per language-pair, for a total of 5,200.

To annotate this data, we developed a context annotation typology (described in section 3.1). Then, we used Stanza (Qi et al., 2020), a POS tagger that uses the Universal Dependencies (UD)[1] framework of 17 tags, to identify the POS patterns associated with context-dependent segments.

Finally, we added 8,000 TUs to the dataset for the same four language varieties, using data from seven different clients within the customer-support domain, but with distinct subject matters (such as technology and retail) in order to test the reliability of our first annotation across a range of content. This time around, we considered only segments with gender-related issues, since these were the most frequent issue type in our dataset. The goal of this analysis was to verify whether our hypothesis would hold across new content; that the POS patterns found would indeed help identify context-dependent segments. In total, we analyzed 13,200 TUs according to these predefined context-dependency patterns.

The number of TUs for all of the experiments is displayed in Table 1.

	N° of TUs
Experiment 1	5,200
Experiment 2	8,000
Total	13,200

Table 1: Distribution of TUs per experiment

3.1 Annotation Typology

Our annotation typology was based on Castilho et al. (2021) and has four categories: gender agreement, number agreement, ellipsis, and terminology. Additionally, we included a fifth category, register, to control the level of formality/informality within a document. For the annotation process, we consider context as all the linguistic information that precedes or follows a segment, and is essential for correct interpretation (Melby and Foster, 2010).

3.1.1 Gender agreement

While English is generally considered a neutral grammatical language, most of the Romance languages (including those analyzed) have rich morphological marking strategies to express gender. We annotated a gender agreement issue wherever it was not possible to disclose the gender of the referent within the segment.

1. Thank you for contacting us.

   a. Gendered TU: **Obrigado** por entrar em contacto connosco.
   b. Gender neutral TU: **Agradeço** por entrar em contacto connosco.

The word *Obrigado* (*thank you*) inflects for gender in Portuguese, meaning that it has both a masculine and a feminine form. Whether the masculine or feminine form is used depends on the gender of the speaker. As such, we annotate this segment for gender agreement. Were we to replace the adjective *Obrigado* with *Agradeço* (a verb with the equivalent meaning) we would remove the gender constraints and generate a context-independent segment.

3.1.2 Number agreement

This category denotes segments that require number agreement between pronouns and their referent. We apply this category where we find a number agreement issue within an intersentential referential chain.

---

[1] https://universaldependencies.org/

2. Context-dependent: John really liked the product. (...) **They** were very high quality.

3. Context-independent: John really liked the product. (...) **It** was very high quality.

Segment (2) shows a number agreement issue, whereas segment (3) uses the pronoun with correct number agreement.

3.1.3 Ellipsis

Ellipsis is a linguistic mechanism defined as the omission of one or more elements within a clause to avoid repetition whilst maintaining textual cohesion. Some languages are more lenient around ellipsis, whereas others are more restrictive. This syntactic imbalance between languages poses challenges for the MT systems that must predict elements that are implied rather than expressed. We use this category for the omission of information within segments, where this omission compromised comprehension.

4. Any time that you make a change in your account, **even if it's a photo**, we will send you an e-mail.

   a. CD - PT: Sempre que efetuas uma alteração na tua conta Ø nós iremos enviar-te um e-mail.

   b. CI - PT: Sempre que efetuas uma alteração na tua conta, **mesmo que seja uma foto**, nós iremos enviar-te um e-mail.

In segment (a), information (in bold) was omitted from the target text, affecting its intended interpretation, whereas in segment (b), the original information is retained.

3.1.4 Terminology

Terminology targets words/terms that constitute a set of vocabulary within a field of knowledge. The polysemous nature of terminology items makes them a source of ambiguity. They diverge from *common* lexical units since they pertain to specialised domains (such as tourism, tech, retail). We applied this annotation wherever a term was wrongly selected due to poor contextual information within the sentence boundaries.

5. Thank you for contacting our **customer support**.

   a. ES: Gracias por ponerse en contacto con nuestro **servicio de Soporte al Client**.

   b. ES: Gracias por ponerse en contacto con nuestro servicio de **Atención al Cliente**.

In this example, the correct translation (b) uses *Atención al Cliente*, as *Atención* is the term stipulated by the client for use in this context.

3.1.5 Register

Register targets aspects resulting from language modulation where speakers adapt their discourse according to the audience, observed through politeness strategies. Formality and informality are, thus, concepts present in most languages. Nevertheless, the way these concepts are expressed vary across languages, which can result in MT inaccuracies. In the annotation process, the category Register should be applied whenever there was a register issue, e.g., use of informality instead of formality and vice versa. See the following examples of formal and informal translations for the same source text.

6. Thank you for **contacting** URL-0.

   a. Formal: Obrigado por **entrar** em contacto com a URL-0.

   b. Informal: Obrigado por **entrares** em contacto com a URL-0.

4 POS patterns distribution

The annotation of the 5,200 TUs in the first experiment showed that 6.5% of the segments were context-dependent. As for the context-related issues, gender agreement corresponded to 98% of all cases and the remaining had residual occurrences, namely register (1.2%), terminology (0.3%), and ellipsis (0.5%).

After the annotation step, we applied Stanza to identify POS patterns. To do so, we used the POS category that was context-dependent, marked with an asterisk (*), and the category that precedes or follows it. As a result, we found eight patterns that are presented in Table 2.

Out of the 8,000 analyzed segments in the second experiment, only 4,224 TUs matched with one of the eight patterns found for both context-dependent and context-independent. Results showed that 19% of these TUs were context-dependent and 81% were context-independent.

As for the eight POS patterns (see Table 2), they were not only common between language variants but also between different languages (e.g. $AUX + *ADJ$).

Language	POS Pattern
PT	$PRON + *VERB$
PT	$*VERB + ADP$
PT-BR	$*VERB$
ES	$ADP + *PRON$
ES-LATAM	
PT	$VERB + *PRON$
PT-BR	$VERB + *ADJ$
ES	$AUX + *ADJ$
ES-LATAM	$AUX + *VERB$

Table 2: Context-dependent POS patterns per language

Language	Context-dependent	Context-independent
PT	8%	92%
PT-BR	27%	73%
ES	18%	82%
ES-LATAM	23%	77%
Total	19%	81%

Table 3: Context-(in)dependent TUs per language

The two experiments showed that, out of 13,200 of the most frequent TUs from various clients, 10% (1,298) were context-dependent. One of our hypothesis was that through POS information it would be possible to automatically identify all segments that were context-dependent. We were able to distinguish two patterns that only occurred in context-dependent segments, namely $ADP + *PRON$ for Spanish (ES) and Spanish Latam (ES-LATAM), and $*VERB$ for Portuguese (PT) and Brazilian Portuguese (PT-BR). The remaining also occurred in context-independent ones, which led us to perform a more fine-grained analysis on context-dependent TUs, as it will be described in Section 5.

4.1 Gender-neutral TUs

From all the context-dependent TUs, 97% were related with gender issues. For the more frequent ones, we proposed gender-neutral alternatives, by replacing the gendered words or expressions with neutral alternatives, as shown in the following examples:

7. EN: Thank you for your patience.

    a. Context-dependent: **Obrigado** pela sua paciência.

    b. Gender neutral: **Agradeço** pela sua paciência.

8. EN: Rest assured, there have been no discrepancies with the rewards.

    a. Context-dependent: Estoy **encantada** de proporcionarte más información hoy.

    b. Gender neutral: **Tengo todo el gusto** de proporcionarte más información hoy.

For TUs such as the one in the example (7), we replaced the participial verb "Obrigado" with a verb with equivalent meaning, however, without any gender constraints, therefore, turning these segments into gender neutral. As for TUs such as the ones in (8), we replaced the gendered adjectives with a nominal expression with an equivalent meaning and gender-neutral, thus preserving the meaning of the original message. All the proposed gender-neutral segments were verified by professional linguists and translators who are native speakers of these languages and were integrated into production, allowing for a more inclusive content.

5 POS patterns and context-dependent words

After the previous experiment, where we identified that the POS patterns were not exclusive of context-dependent segments, also occurring in context-independent ones, we performed a root-cause analysis to understand if there were distinctive features. While annotating, it was clear that context-dependent segments usually involved specific keywords that were also very frequent amongst all the data, such as *"obrigado(a)"* for PT or *"entantado(a)"* for ES. Therefore, a new analysis was conducted, in which we aimed to verify if using these frequent keywords in conjunction with the POS patterns would facilitate the detection of context-dependent segments.

5.1 Methodology

We conducted two new experiments, one in customer-support emails and other in customer-support chat messages. Firstly, using the previously analyzed data we gathered a list of all of the keywords that triggered context-related issues, exclusively associated with gender. Secondly, we

analyzed a new dataset with 2,000 TUs from 5 clients (500 units per language for PT, PT-BR, ES, and ES-LATAM) for emails, and another dataset of 1,052 chat messages segments only for PT-BR. This step involved the classification of all the segments by searching the eight POS patterns. The final task was to search for the frequent keywords in the TUs that matched with one of the POS patterns and verify if the segments were in fact context-dependent.

5.2 Results

For the emails dataset, results show that 54.6% of the TUs matched with one of the eight POS patterns. The TUs that matched with both a POS pattern and the keywords for each language accounted for 14% of the cases (Table 4). Our results show too that segments that matched with a POS pattern and a keyword were always considered context-dependent.

For PT, 4.5% of cases matched with both a pattern and at least one keyword, thus were context-dependent. For PT-BR, the same occurred in 5.2% of the segments. For these language varieties, the results for the keywords and patterns were very similar. *Obrigado(a)* occurring with the patterns $*VERB + ADP$ and $*VERB$ was the most frequent word in these languages (see example below):

Language	Context (in)dependent	Context-dependent
	POS	POS + Keywords
PT	284	49 (4.5%)
PT-BR	287	57 (5.2%)
ES	275	23 (2.1%)
ES-LATAM	246	19 (1.7%)
Total	1,092 (54.6%)	148 (13.6%)

Table 4: POS matches and POS + keywords matches

9. **Thanks** for your time and cooperation.
   Obrigado pelo seu tempo e cooperação.

2.1% of ES TUs and 1.7% of ES-LATAM TUs matched with both the POS patterns and keywords and were classed as context-dependent. Again, results are similar for both language varieties,

since the keywords are also similar. The pronoun *nosotros* that occurred with the pattern $ADP + *PRON$, and the adjectives *encantado*, and *satisfecho*, that occurred with the pattern $AUX + *ADJ$, were the most frequent words:

10. That's no problem at all, I'm **delighted** everything is sorted.
    Eso no es ningún problema, estoy **encantado** de que todo esté resuelto.

For the final experiment with chat messages data, results showed that 43.5% of the segments matched with one of the POS patterns and 32.1% were context-dependent. Similar to the results for PT-BR, *obrigado(a)* was the most frequent keyword.

A final evaluation was performed in order to verify that all results were context-dependent. This confirmed all aforementioned results were in fact context-dependent.

The results from this experiment showed that the keywords seem to be effective as a disambiguation step.

6 Conclusions

We analyzed 13,200 TUs and identified over 1,298 context-dependent segments of which 1,263 had gender constraints. For these gender-constrained segments, we proposed gender-neutral alternatives by either replacing the gendered words with neutral alternatives or by syntactically manipulating the sentence in order to obtain a gender-neutral sentence with an equivalent meaning.

We hypothesized that POS information would enable us to automatically identify all context-dependent segments. In addressing this, we were able to identify eight patterns: one exclusive for PT (*PRON+VERB), two exclusives for PT and PT-BR (*VERB and *VERB+ADP), one for ES and ES-LATAM (ADP+*PRON) and the remaining were common for all languages (VERB+*ADJ, AUX+*ADJ, VERB+*PRON and AUX+*VERB). However, our experiments with POS lead us to conclude that the POS patterns do not discriminate sufficiently between context-dependent and context-independent classification in all cases.

Conducting a root-cause analysis, we notice that context-dependent segments involve specific words. For instance, the 3rd person singular pronouns for PT and PT-BR (-lo and -la) and 1st person plural pronouns for ES and ES-LATAM

(*nosotros*) were very common and only knowing that these pronouns can occur is already very informative. We also analyzed adjectives such as *satisfeito(a)*, *interessado(a)* or *encantado(a)* and *emocionado(a)*, and other similar adjectives that allow one to express appreciation or dissatisfaction and also specific participial verb forms such as *obrigado/obrigada*. The frequency of this specific vocabulary in the customer-support domain may in fact tell us more than the POS information. Another contribution of this work was to verify that these very frequent lemmas combined with the POS information can add value to the improvements of a system detecting context-(in)dependent TUs and generating gender-neutral alternatives when possible.

Following our evaluation, we were able to suggest gender-neutral TUs for the most frequent segments. All the segments proposed were then verified by professional linguists and translators who are native speakers of the languages in question. After review, the segments were injected into our production TM database. These segments maintain equivalent meaning but are now no longer context-dependent.

The work described is applied in production at Unbabel, aligned with clients' reports and NLP modules. Since completing work to transform the most frequent TUs analyzed into their gender-neutral equivalents, we have seen a reduction in gender-agreement errors (since our translations are more consistent at the document level), and a reduction in editing time. Although we demonstrate this approach in the context of the customer-support domain, we argue that it can be applied elsewhere: to other domains, industries and NLP pipelines. This work is now being used as validation and test sets for Large Language Models assessment for gender-neutral MT.

7 Acknowledgements

References

Bawden, Rachel, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2017. Evaluating discourse phenomena in neural machine translation. *arXiv preprint arXiv:1711.00513*.

Bawden, Rachel. 2018. *Going beyond the sentence: Contextual machine translation of dialogue*. Ph.D. thesis, Université Paris-Saclay (ComUE).

Bowker, Lynne and Des Fisher. 2010. Computer-aided translation. *Handbook of translation studies*, 1:60–65.

Bublitz, Wolfram. 2011. Cohesion and coherence. *Discursive pragmatics*, 8:37–49.

Cai, Xinyi and Deyi Xiong. 2020. A test suite for evaluating discourse phenomena in document-level neural machine translation. In *Proceedings of the Second International Workshop of Discourse Processing*, pages 13–17.

Castilho, Sheila, João Lucas Cavalheiro Camargo, Miguel Menezes, and Andy Way. 2021. Dela corpus-a document-level corpus annotated with context-related issues. Association for Computational Linguistics (ACL).

Gehring, Jonas, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Guillou, Liane, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the english–german mt systems at wmt 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577.

Hieber, Felix, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*.

House, Juliane. 2006. Text and context in translation. *Journal of pragmatics*, 38(3):338–358.

Jwalapuram, Prathyusha, Barbara Rychalska, Shafiq Joty, and Dominika Basaj. 2020. Can your context-aware mt system pass the dip benchmark tests?: Evaluation benchmarks for discourse phenomena in machine translation. *arXiv preprint arXiv:2004.14607*.

Koehn, Philipp. 2020. *Neural machine translation*. Cambridge University Press.

Lopes, António V, M Amin Farajian, Rachel Bawden, Michael Zhang, and André FT Martins. 2020. Document-level neural mt: A systematic comparison. In *22nd Annual Conference of the European Association for Machine Translation*, pages 225–234.

Melby, Alan K and Christopher Foster. 2010. Context in translation: Definition, access and teamwork. *Translation & Interpreting, The*, 2(2):1–15.

Müller, Mathias, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. *arXiv preprint arXiv:1810.02268*.

Piergentili, Andrea, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. From inclusive language to gender-neutral machine translation. *arXiv preprint arXiv:2301.10075*.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the ACL 2020, Annual Meeting of the Association for Computational Linguistics*, pages 101–108.

Sun, Tony, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english. *arXiv preprint arXiv:2102.06788*.

Voita, Elena, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. *arXiv preprint arXiv:1805.10163*.

Yin, Kayo, Patrick Fernandes, André FT Martins, and Graham Neubig. 2021a. When does translation require context? a data-driven, multilingual exploration. *arXiv preprint arXiv:2109.07446*.

Yin, Kayo, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André FT Martins, and Graham Neubig. 2021b. Do context-aware translation models pay the right attention? *arXiv preprint arXiv:2105.06977*.