Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3017–3025
Marseille, 20-25 June 2022

How Much Context Span is Enough? Examining Context-Related Issues for Document-level MT

Sheila Castilho
Dublin City University
ADAPT Centre/School of Computing
sheila.castilho@adaptcentre.ie

Abstract

This paper analyses how much context span is necessary to solve different context-related issues, namely, reference, ellipsis, gender, number, lexical ambiguity, and terminology when translating from English into Portuguese. We use the DELA corpus, which consists of 60 documents and six different domains (subtitles, literary, news, reviews, medical, and legislation). We find that the shortest context span to disambiguate issues can appear in different positions in the document including preceding, following, global, world knowledge; and that the average length depends on the issue types as well as the domain. Additionally, we show that the standard approach of relying on only two preceding sentences as context might not be enough depending on the domain and issue types.

Keywords: document-level, context span, machine translation

1. Introduction

Recently, there has been a rise in Neural Machine Translation (NMT) research, and NMT is now widely used in a variety of fields, mainly due to advancements in neural models (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). Consequently, increasing efforts have been made in order to include discourse into these systems (Wang, 2019; Lopes et al., 2020).

Despite the fact that *context* is extensively used in translation and interpreting literature (Baker, 2006), it still lacks a specific definition for practical purposes, such as in a professional translator's day-to-day work (Melby and Foster, 2010). For Melby and Foster (2010, p.3), context can be categorised into *non-text* (non-linguistic variables) and *text* (linguistic aspects).

In this work, we examine the context span necessary to solve six different context-related issues, namely, reference, ellipsis, gender, number, lexical ambiguity, and terminology. We adopt Melby and Foster (2010)'s view of context that is important to the analysis of translations, and focus (i) on the *co-text*, i.e. the boundaries within the document translated, and (ii) in the *non-text*, where the name of the authors, speakers, and products have an effect on the translation. We use the DELA corpus (Castilho et al., 2021), a document-level corpus annotated with the aforementioned context-aware issues when translating from English (EN) into Brazilian-Portuguese (PT-BR). It consists of 60 documents and six different domains (subtitles, literary, news, reviews, medical, and legislation). We find that the shortest context span to solve the issues can appear in various positions in the document including preceding, following, global, and world knowledge contexts. Moreover, we find that the average context length depends on the issue types as well as the domain.

2. Related Work

Even though the Machine Translation (MT) community has become more interested in discourse MT, the definition of what constitutes a document-level MT is still unclear (Castilho et al., 2020). Furthermore, it is also unclear whether document-level NMT models "rely on the 'right' context that is actually sufficient to disambiguate difficult translations" (Yin et al., 2021, p.788).

Few attempts have been made in order to tackle the question of how much context span is needed for document-level MT. Castilho et al. (2020) tested the context span for the translation of 300 sentences in three different domains (reviews, subtitles, and literature) and showed that over 33% of the sentences tested required more context than the sentence itself to be translated or evaluated, and from those, 23% required more than two previous sentences to be properly evaluated. An interesting finding is that ambiguity, terminology, and gender agreement were the most common issues to hinder translation, and moreover, there were observable differences in issues and context span between domains. In line with these results, Rikters and Nakazawa (2021) have also observed that 20% of the antecedents of anaphoras of an English corpus appear more than two sentences before the current sentence where the anaphoras appear. This means that existing document-level MT models which consider only 1-2 previous sentences are not sufficient.

Yin et al. (2021) investigate what context is intrinsically useful to disambiguate translation phenomena, namely pronoun anaphora ('it' and 'they'), and word sense (automatically identified). The authors had 20 professional translators annotating 400 examples of contrastive translation between English and French, with 5 different context levels: no context, previous source sentence only, previous target sentence only, previous source sentence and target sentence, and the

5 previous source and target sentences. Their results show that for the anaphora task, when inter-sentential context is available, translators make use of the context but selected more supporting context from the target side, while for the word sense, inter-sentential context is seldom highlighted and more annotations are performed on the source sentences. The authors hypothesise that translators might pay more attention to the source sentences to understand the source material when translating, however, during disambiguation translators rely more often on the target sentences.

Adding a broader context when assessing the quality of MT systems has also been attempted in MT evaluation. One current way of evaluating document-level issues is the use of contrastive test suites, and, even though several have been buitl (Bawden et al., 2018; Guillou et al., 2018; Müller et al., 2018; Voita et al., 2019; Cai and Xiong, 2020), test suites with document-level boundaries are still scarce (Vojtěchová et al., 2019; Rysová et al., 2019) and none of them examine the actual context span.

Although not looking into the context span, Castilho (2020; 2021) tested for the differences in inter-annotator agreement (IAA) between single sentence and document-level set-ups. The results of both studies support the need to add context to MT evaluation as they showed that methodologies in which translators assign one score per sentence in context avoids mis-evaluation cases - extremely common in the random sentences-based evaluation set-ups. Furthermore, the author concludes that the misevaluation issue is especially problematic when assessing the quality of NMT systems as they have an improved fluency level and therefore, single random sentence evaluation method should be avoided.

As can be seen, there is a need for the implementation of context-aware evaluation in MT, and with that, the need for more research on the definition of document-level and the context span necessary to solve ambiguities. In this paper, we attempt to shed some light on this issue by examining the context span necessary to solve context-related issues annotated in the DELA corpus.

3. Identifying the Context Span for Context-Related Issues

3.1. Context-Related Issues

In this work, we use the DELA corpus, a document-level corpus described in (Castilho et al., 2021) where context-related issues were annotated. The corpus contains 60 full documents and was compiled with six different domains: subtitles (9 docs), literary (4 docs), news (15 docs), reviews (28 docs), medical (3 docs), and legislation (1 doc), with a total of 3710 sentences, and an average of 15.57 words per sentence. We use the EN part of the corpus with the annotated issues and counted how many preceding or following sentences were necessary in order to solve each issue.

The DELA corpus was annotated by three annotators with six context-aware issues that are challenging for MT when translating from EN into PT-BR to be annotated, namely reference, ellipsis, gender, number, lexical ambiguity, and terminology. Only issues that could not be solved *within* the sentence were annotated. We briefly re-visit the description of issues here:

Reference was annotated whenever there was a disruption or ambiguity in the referential chain, e.g.:
It is understandable though since ***it*** *was shipped from China.*
reference → it = the ship

Ellipsis was annotated when the omission of information affects the translation of that specific single sentence, e.g.:
In my laughter, I bellied out a "YES, I do!!"
ellipsis → do = think

Gender was annotated whenever a gender ambiguity was found to be unsolvable within the sentence itself, e.g.:
I'm surprised to see you back so early.
gender → surprised = feminine

Number was annotated whenever a number ambiguity was found within the referential chain, such as (i) noun or pronoun, (ii) verb and noun/pronoun, (iii) adjective, caused by lack of enough contextual information, e.g.:
I was praying for you.
number → you = plural

Lexical Ambiguity was annotated whenever a word or a phrase appeared to be detrimental to the translation and understandable only within the broader context, e.g.:
He came back in the house and said "So you think this is funny?!" up the stairway at me and I LOST IT.
lexical ambiguity → lose something *vs* to lose control

Terminology was annotated when a wrongly domain-specific word translation caused by contextual poor sentences was found, e.g.:
The center will also conduct testing (power curve, mechanical loads, noise, and power quality) at its own experimental wind farm
terminology → generalised lexic (farm) *vs* domain-specific lexicon (park)

3.2. Types of Context Span

Considering the DELA corpus annotation, we looked into the *shortest* context span necessary to solve every issue annotated in the EN part of the corpus. We have categorised the context span into:

- Preceding (PREC): the shortest context span consists only of immediate sentences BEFORE the source sentence.
- Following (FOLL): the shortest context span consists only of immediate sentences AFTER the source sentence.

Full Corpus	PREC	FOLL	Prec+Foll	Prec/Foll	GLOB	W	TOTAL	%
Reference	201	24	0	5	2	0	232	17.14
Ellipsis	27	9	1	1	1	0	39	2.88
Gender	348	121	5	5	14	2	495	36.58
Number	116	25	6	2	0	0	149	11.01
Lexical Ambiguity	212	121	22	6	56	9	426	31.48
Terminology	1	0	0	0	7	4	12	0.88
TOTAL	905	300	34	19	80	15	1353	-
%	66.88	22.17	2.51	1.40	5.91	1.1	-	

Table 1: Total number of issues found in the corpus and their respective location regarding the source sentence they were found.

- Preceding + Following (Prec+Foll): the shortest context span consists of immediate sentences before AND after the source sentence.
- Preceding / Following (Prec/Foll): the shortest context span consists of immediate sentences EITHER before OR after the source sentence.
- Global (GLOB): the context span required does not lie in a single sentence (or a chunk of few sentences), therefore, the full text is needed in order to solve the issue.
- World (W): the context span required does not lie in the full text as it crosses the document boundaries.

4. Results

This section presents the results found when examining the DELA corpus in terms of i) context position (4.1) which is the position of the sentence(s) that solves the ambiguity of the tagged issues, i.e. preceding, following, etc.; and ii) context span (4.2) which refers to the amount of sentences necessary to give the minimum amount of context for those issues to be solved, that is, the length of the context.

4.1. Context Position

In this section, we look at the number of issues found in the corpus and the position of the sentence(s) that were used to solve the issues tagged, i.e. PREC, FOLL, GLOB, etc. Table 1 shows the total number of issues found in the whole corpus, along with the position of the context span necessary to solve the issues.

As can be seen, the majority of the issues have their shortest context span necessary to be solved positioned before (PREC) the source sentence (66%), followed by a context span after (FOLL) the source sentence (22%). We note that the most common types of issues annotated in the corpus are gender (36%), lexical ambiguity (31%), reference (17%), and number (11%). Interestingly, while most of the gender issues tagged can be solved with a previous context span (over 70%) - even though a great number of following context (24%) and a few global (2%) can be observed - the context position for lexical ambiguity type of issue is more diverse, with the highest being preceding context (49%) but with a great number of issues also being solved with following (28%), preceding+following (5%), global (13%), and world (2%) context spans. As expected due to its nature (see section 3.1), the reference issue is mostly solved with preceding context (86%). The context span necessary to solve the grammatical number issues is also mostly positioned preceding (77%) the issue, which we believe is due to the type of documents used in the subtitles domain which had a higher number of grammatical number issue (see sections 4.1.4 and 4.2.4). In order to see if any definite patterns can be observed for different domains, we have a more in-depth look at each one of them in Table 2.

4.1.1. Literary

In table 2, we observe that in the literary corpus, lexical ambiguity is the most common type of context-related issue annotated (51%), followed by gender (35%) and reference (7%). While lexical ambiguity presents a diverse context span position, the context span for gender is mostly found in preceding context (even though it shows a great number of following context as well). It is worth noticing that there are no terminology issues annotated in the literary corpus, which might be expected as the excerpts and books chosen do not contain any technical language.

4.1.2. Review

It is possible to see that gender is the most common type of issue annotated (42%), followed by reference (30%), and lexical ambiguity (22%) issues (Table 2). The majority of the issues has their shortest context position preceding the source sentence (176 issues, 63%). This seems to be a characteristic of the domain since, in reviews, it is quite normal for the user to write the name of the product/place once (or not at all) and then just refer to the product/place as ‘it’ or ‘they’, which are pronouns that can lead to these types of ambiguity issues (see section 3.1). It is worth noticing that, in this domain, a great number of issues has their shortest context span position following the source sentence (23%), and 19% of the times (12 instances), lexical ambiguity issues needed the global context of the review to be solved.

LITERARY	PREC	FOLL	Prec+Foll	Prec/Foll	GLOB	W	TOTAL	%
Reference	23	2	0	0	0	0	25	7.69
Ellipsis	10	4	0	0	0	0	14	4.31
Gender	70	37	2	0	6	0	115	35.38
Number	3	1	0	1	0	0	5	1.54
Lexical Ambiguity	85	55	0	2	22	2	166	51.08
Terminology	0	0	0	0	0	0	0	0
TOTAL	191	99	2	3	28	2	325	-
%	58.77	30.46	0.62	0.92	8.62	0.66	-	
REVIEW	PREC	FOLL	Prec+Foll	Prec/Foll	GLOB	W	TOTAL	%
Reference	64	15	0	4	1	0	84	30.43
Ellipsis	5	3	0	1	0	0	9	3.26
Gender	79	27	0	5	6	0	117	42.39
Number	2	0	0	0	0	0	2	0.72
Lexical Ambiguity	25	21	1	4	12	0	63	22.83
Terminology	1	0	0	0	0	0	1	0.36
TOTAL	176	66	1	14	19	0	276	-
%	63.77	23.91	0.36	5.07	6.88	0.00	-	
NEWS	PREC	FOLL	Prec+Foll	Prec/Foll	GLOB	W	TOTAL	%
Reference	15	3	0	0	1	0	19	6.93
Ellipsis	3	2	1	0	0	0	6	2.19
Gender	73	44	3	0	2	1	123	44.89
Number	2	0	6	0	0	0	8	2.92
Lexical Ambiguity	50	35	8	0	11	7	111	40.51
Terminology	0	0	0	0	3	4	7	2.55
TOTAL	143	84	18	0	17	12	274	
%	52.19	30.66	6.57	0.00	6.20	4.38	-	
SUBTITLES	PREC	FOLL	Prec+Foll	Prec/Foll	GLOB	W	TOTAL	%
Reference	88	3	0	1	0	0	92	23.23
Ellipsis	7	0	0	0	0	0	7	1.77
Gender	96	10	0	0	0	1	107	27.02
Number	103	24	0	1	0	0	128	32.32
Lexical Ambiguity	36	6	13	0	5	0	60	15.15
Terminology	0	0	0	0	2	0	2	0.51
TOTAL	330	43	13	2	7	1	396	-
%	83.33	10.86	3.28	0.51	1.77	0.25	-	
MEDICAL	PREC	FOLL	Prec+Foll	Prec/Foll	GLOB	W	TOTAL	%
Reference	1	0	0	0	0	0	1	2.33
Ellipsis	1	0	0	0	1	0	2	4.65
Gender	14	1	0	0	0	0	15	34.88
Number	0	0	0	0	0	0	0	0
Lexical Ambiguity	13	4	0	0	6	0	23	53.49
Terminology	0	0	0	0	2	0	2	4.65
TOTAL	29	5	0	0	9	0	43	
%	67.44	11.63	0.00	0.00	20.93	0.00	-	
LEGISLATION	PREC	FOLL	Prec+Foll	Prec/Foll	GLOB	W	TOTAL	%
Reference	10	1	0	0	0	0	11	28.21
Ellipsis	1	0	0	0	0	0	1	2.56
Gender	16	2	0	0	0	0	18	46.15
Number	6	0	0	0	0	0	6	15.38
Lexical Ambiguity	3	0	0	0	0	0	3	7.69
Terminology	0	0	0	0	0	0	0	0
TOTAL	36	3	0	0	0	0	39	-
%	92.31	7.69	0.00	0.00	0.00	0.00	-	

Table 2: Total number of issues found in every domain and their respective location regarding the source sentence they were found.

Full Corpus	PREC		FOLL		Av. Span*	
	avg	median	avg	median	avg	median
Reference	2.62	2.00	1.80	1.00	2.53	1.25
Ellipsis	2.65	1.00	0.60	0.50	2.07	1.00
Gender	10.62	2.00	4.93	2.25	9.42	2.00
Number	6.85	1.50	2.28	0.50	7.07	2.00
Lexical Ambiguity	12.16	1.50	5.06	1.78	9.37	1.00
Terminology	0.17	2.00	0.00	0.00	0.00*	0.00*
OVERALL	11.39	2.00	5.37	2.00	9.69	2.00

Table 3: Context span (in sentences) for the full corpus domain for each preceding and following context. Note that A. Span* relates to the span when both preceding and following context spans are considered. *We note for terminology, most of the issues were found to have a global context, with just one issue being found to have a preceding context.

4.1.3. News

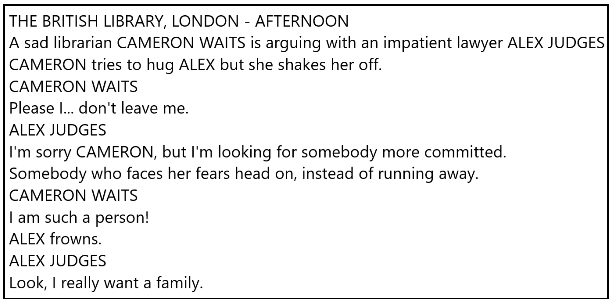
Similar to the review domain, the news corpus also has gender as the most common issue annotated (44%), followed very closely by lexical ambiguity (40%) (Table 2). Following the trend, issues have their shortest context span preceding the source sentence (52%) but a great number still shows the shortest context span after the source (30%). We note again that, for the lexical ambiguity issues, context span positions are more diverse than for the other issues. Finally, terminology issues have their context span located globally, or world knowledge is needed.

4.1.4. Subtitles

We observe that differently from the other domains, the subtitle corpus has the grammatical number issues as the most annotated issues (32%), closely followed by gender (27%), and reference (23%) (Table 2). This is very interesting and shows a very specific characteristic of the domain which, by being made up completely of subtitles from TED Talks, it has the use of the general ‘you’ when the speaker refers to the audience. In most cases, when translated into Portuguese, this ‘you’ needs to be translated as the plural form. We note again that the majority of the issues have their shortest context position preceding the source sentence (total of 83%) but both gender and number issues also present a high FOLL (following the source sentence) context position (10 issues or 9%, and 24 issues or 18% respectively). Lexical ambiguity has also showed a diverse position for the context span in this domain.

4.1.5. Medical

Because of its small size, the medical domain contained few issues, where the most annotated is lexical ambiguity (53%) followed by gender (34%) (Table 2). The same pattern about context position can be observed, where the shortest context span is mostly made up by preceding sentences (67%), but a few of the issues also present following context (11%). Lexical ambiguity, once again, has the context span position diverse.



THE BRITISH LIBRARY, LONDON - AFTERNOON
A sad librarian CAMERON WAITS is arguing with an impatient lawyer ALEX JUDGES
CAMERON tries to hug ALEX but she shakes her off.
CAMERON WAITS
Please I... don't leave me.
ALEX JUDGES
I'm sorry CAMERON, but I'm looking for somebody more committed.
Somebody who faces her fears head on, instead of running away.
CAMERON WAITS
I am such a person!
ALEX frowns.
ALEX JUDGES
Look, I really want a family.

Figure 1: Example of the play script format included in the literary domain. Note that the name of every character is repeated ever time they speak.

4.1.6. Legislation

Similar to the medical corpus (Table 2), the corpus with the speeches from the European parliament also presented a few number of issues, where the most tagged ones were gender (46%), reference (28%) and number (15%). The majority of issues had the shortest context span position before the source sentence (92%).

4.2. Context Span

In this section, we examine the context span, which is the length needed to solve the issues tagged in the DELA corpus. Table 3 shows the statistics for the whole corpus, including the average and the median. We note that the average preceding context span tends to be longer than the average following context span (11.39 and 5.37 sentences long respectively) for all the issues, with the same median. Gender and lexical ambiguity have the longest average context span followed by the grammatical number (when all context span types are considered together, i.e PREC, FOLL, PREC+FOLL, and PREC/FOLL). But we note that the median for Gender (2) is longer than for Lexical ambiguity (1). Ellipsis is the one with the shortest context spans when looking at the averages. For a more in-depth analysis of the context span, we also look into each domain separately in Table 4.

LIT	PREC		FOLL		Av. Span*	
	Av. lgth	Long.	Av. lgth	Long.	avg	med
Ref.	4.3	25	1	1	4.04	1.2
Ellip.	2.7	8	1	1	2.2	1.0
Gend.	29.4	295	15.7	105	24.6	2.0
Numb.	1.4	2	1	1	1.3	2.0
L.Amb.	59.1	435	18.6	74	42.8	1.0
Term.	0	0	0	0	0	0
O.ALL	38.37	-	15.77	-	30.5	7.0
REV	PREC		FOLL		Av. Span*	
	Av. lgth	Long.	Av. lgth	Long.	avg	med
Ref.	4.0	19	3.6	13	3.9	2.0
Ellip.	1.4	2	1.2	2	1.3	1.0
Gend.	4.2	25	4.0	13	4.1	2.0
Numb.	1	1	0	0	0	0
L.Amb.	2.0	8	3.8	16	3.0	1.0
Term.	1	1	0	0	0	0
O.ALL	3.7	-	3.7	-	3.7	2.0
NEWS	PREC		FOLL		Av. Span*	
	Av. lgth	Long.	Av. lgth	Long.	avg	med
Ref.	1.6	3	1.3	1	1.6	1.5
Ellip.	8.2	29	1.3	2	5.2	1.0
Gend.	7.7	33	5.3	23	6.8	2.0
Numb.	2.1	3	1.5	2	1.8	2.0
L.Amb.	4.7	32	3.1	16	4.0	2.0
Term.	0	-	0	-	0	0
O.ALL	6.0	-	4.0	-	5.2	2.0
SUBS	PREC		FOLL		Av. Span*	
	Av. lgth	Long.	Av. lgth	Long.	avg	med
Ref.	2.9	30	1	1	2.8	1.0
Ellip.	1.5	5	0	0	1.5	1.0
Gend.	17	171	2	5	15.5	2.0
Numb.	33.9	143	11.1	23	29.5	17.0
L.Amb.	3.2	41	2.6	11	3.0	1.0
Term.	0	0	0	0	0	0
O.ALL	16.1	-	6.1	-	14.7	3.0
MED	PREC		FOLL		Av. Span*	
	Av. lgth	Long.	Av. lgth	Long.	avg	med
Ref.	1	1	0	0	1	1.0
Ellip.	1	1	0	0	1	1.0
Gend.	2.5	6	1	1	2.4	2.0
Numb.	0	0	0	0	0	0
L.Amb.	1.25	2	1	1	1.1	1.0
Term.	0	0	0	0	0	0
O.ALL	1.8	-	1	-	1.7	1.0
LEG	PREC		FOLL		Av. Span*	
	Av. lgth	Long.	Av. lgth	Long.	avg	med
Ref.	1.8	6	2	2	1.8	1.0
Ellip.	1	1	0	0	1	1.0
Gend.	2.8	9	1.5	2	2.9	1.0
Numb.	2.6	5	0	0	2.6	2.5
L.Amb.	2.5	4	1	1	2	1.0
Term.	0	0	0	0	0	0
O.ALL	2.2	-	1.5	-	2.14	1.0

Table 4: Context span length (in sentences) in all domains and their longest span for each preceding and following context. Note that Av. Span* in the last column relates to the span when both preceding and following context spans are considered.

LIT	PREC		FOLL		Av. Span*	
	Av. lgth	Long.	Av. lgth	Long.	avg	med
Reference	**2.8**	**13**	1	1	**2.69**	1.0
Ellipsis	2.7	8	1	1	2.2	1.0
Gender	**25.4**	103	**18.2**	105	**23.2**	7.5
Number	1.4	2	1	1	1.3	1.0
Lex. Amb.	**10.5**	**31**	**9.95**	74	**10.2**	2.5
Terminology	0	0	0	0	0	0
OVERALL	**8.7**	-	**9.2**	-	**8.98**	3.0

Table 5: Context span length (in sentences) in the literature domain when the play script document is not considered. The numbers in **bold** refer to differences between the two calculations (see Table 4).

4.2.1. Literary

The context span for the literary corpus is, on average, longer for preceding (38.37 sentences long) than for following context (15.77), as shown in Table 4. Lexical ambiguity presents the longest average context span (42), followed by gender (24.6), with the median been higher for gender (2). One reason for the context span to be lengthy for this domain is because one of the documents in the corpus is a *play script* whose format is very specific as it displays the name of the character every time they speak (see figure 1). For that reason, we also calculate the average context span without the play script in order to have an idea how much the context span might change. Results in Table 5 show that the context span average for the reference issue for the preceding context decreased (2.83 average, 1 median), and the average for both preceding and following context lengths is down from 4.04 to 2.69 sentences long. For the lexical ambiguity issue, the difference is much larger, where the preceding average context span is 10.57 (down from 59.19), the following average context span is 9.95 (down from 18.63), and the average for both is 10.2 (down from 42.86, with an increase in the median to 2.5). This is interesting as it shows that, on average, the following context span is slightly longer than the preceding one for all the issues (8.78 for PREC and 9.20 for FOLL) making the average context span for the domain 8.98 sentences long. This brings the gender issue as the one needing the longest context span (23.2 average, 7.5 median) on average and lexical ambiguity the second (10.2 average, 2.5 median).

4.2.2. Reviews

Table 4 shows the results for the context length for the review domain. We observe that the lengths for preceding and following context are, on average, very close to each other. But interestingly, lexical ambiguity shows a longer following context span (PREC 2.03, FOLL 3.88). The context length for the gender issues is the longest, followed by reference and lexical ambiguity, where the average context span where all issues and context span type is considered is 3.72 sentences long. We note that the median are the same for reference and gender.

4.2.3. News

Similar to the review domain, the news domain (Table 4) also presents the longest context span for gender (6.84), followed by ellipsis (5.28) and lexical ambiguity (4.09). However, differently, we note that the preceding context span is, on average, longer (6.02) than the following (4.05) context span, and moreover, the general context span for the news domain is longer with 5.25 sentences long on average, with the same median.

4.2.4. Subtitles

In the subtitles domain, the preceding context span is much longer (16.19) than the following span (6.10), as showed in Table 4. Interestingly, the grammatical number issue has the longest span (29.5 average, 17 median) which correlates with the amount of issues relating to grammatical number found in the domain (see Table 2). We note that, not only is grammatical number a very common issue in this domain, but the length of context to solve it is also quite lengthy.

4.2.5. Medical

As seen in the previous section, the medical domain had only a few issues tagged (see Table 4). When looking at the context length to solve those issues, we note that the average of the span is quite short (1.77 sentences in either direction), being the longest span seen for gender in preceding context (2.5 average, 2 median).

4.2.6. Legislation

Similar to the medical domain, the legislation domain also presents a short average context span of 2.14 sentences to either direction. Gender (2.94), number (2.66) and lexical ambiguity (2.00) have the lengthiest average context (Table 4). Interestingly, the reference issue which was the second most annotated issue in this domain (11 in total, just under gender with 18 issues - see table 2) does not need a lengthy span to be solved.

5. Discussion and Conclusions

In this paper, we shed some light on the issue of the definition of document-level, and the context span necessary to solve ambiguities. We used the context-related issues annotated in the DELA corpus, when translating from EN into PT-BR, namely, reference, ellipsis, gender, number, lexical ambiguity, and terminology and examined the *shortest* context span necessary to solve them, and categorise the types of contexts according to their position. We reported the i) Context Position (Section 4.1), and ii) Context Length (Section 4.2). Table 6 shows the summary for context-related issues found, their percentage, the shortest position of the context in relation to the sentence where the issue was tagged, and their average and median length of the context span when looking at the whole corpus. Table 7 shows a break-down by domain regarding most tagged issue, and the average and median context span within the domain.

% of issues	Issue	Position	Av. Length	Median
36.58	Gender	PREC, FOLL	9.42	2.00
31.48	Lex. ambiguity	PREC, Diverse	9.37	1.00
17.14	Reference	PREC	2.53	1.25
11.01	Number	PREC	7.07	2.00
2.88	Ellipsis	PREC	2.07	1.00
0.88	Terminology	Global	-	-

Table 6: Summary of percentage of issues, context span position, and context length found in the DELA corpus.

Regarding the *gender* issue, we found that it was the most tagged issued in the whole corpus (36% of all issues), appearing as one of the most tagged issues in every domain (Tables 6 and 7). This is expected since Portuguese is a language in which grammatical gender (feminine and masculine) plays a very significant role, where word classes like adjectives, articles or pronouns are bound to respect and reflect a word's gender (Grosjean et al., 1994). The context position to solve the gender issue was mostly found in preceding sentences, but with a great number of them being solved with following context. Moreover, gender was also the issue that needed the most amount of context to be resolved on average in the whole corpus (9.42 sentences long - even though the median is the same as grammatical number) and within domains apart from the subtitles domain, in which gender is the second longest. This result indicates that, for languages that observe grammatical gender, this issues could also be a recurrent one and needs a lengthy context to be solved.

Lexical ambiguity was the second most tagged issue (31%) and the amount of context needed to solve it was on average with 9.37 sentences long, with a 1 sentence

Domains	Most tagged	AV Length	Median
Literary	Lex. Ambiguity	10.2	2.50
	Gender	23.2	7.50
	Reference	2.69	1.00
Review	Gender	4.16	2.00
	Reference	3.93	2.00
	Lex. Ambiguity	3.03	2.00
News	Gender	6.84	2.00
	Lex. Ambiguity	4.09	2.00
	Reference	1.61	1.50
Subs	Number	29.5	17.00
	Gender	15.58	2.00
	Reference	2.8	1.00
Medical	Lex. Ambiguity	1.18	1.00
	Gender	2.4	2.00
Legislation	Gender	2.94	1.00
	Reference	1.81	1.00
	Number	2.66	2.50

Table 7: Summary of context length found in the DELA corpus.

median (Table 6). One interesting characteristic of this issue is that it differs from the others in terms of presenting a diverse context position throughout the corpus, with many instances of PREC+FOLL, GLOB and W contexts. Within domains, lexical ambiguity was the most tagged in the literary and medical domains, being also one of the three most tagged issues in the review and news. Interestingly, lexical ambiguity is not the issue that needs the longest context to be solved within the domains, and moreover, it does not appear as one of the three most tagged issues in subtitles and legislation domains (Table 7).

Regarding *reference*, we found that it was the third most annotated issues in the whole corpus, with the majority of the context span as preceding sentences (Table 6). This is not so surprising because of the nature of references, where generally the anaphoric reference is generally first mentioned and then referred to at a later stage. Reference also showed one of the shortest average context length (2.53), with a median of 1.25 sentences, which, we believe is because of the types of documents in the corpus, especially *reviews* and *news* - which had the most numbers of reference issues tagged - as they tend to be shorter documents than the other domains (Table 7) where the referential unit issue tagged is generally not very far from the anaphoric reference.

Number, was the fourth most annotated issue in the corpus, with the majority of the context span as preceding sentences (Table 6), which we believe might be a characteristic of the types of documents in the corpus. Since the *subtitles* domain contained the highest amount of grammatical number issues annotated (Table 7) - which has the use of the general 'you' when the speaker refers to the audience, that needs to be translated as the plural form - the solution for the grammatical number would frequently be before the repeated mentions of the 'you'. But differently from reference, the grammatical number issue needed on average a longer context span needed to solved (7.07 sentences long on average, with a median of 2 sentences). Again, we believe that the type of documents in the corpus played a role here, since the documents in the subtitles domain are longer than reviews and news, therefore, the mention of 'you' was farther in the document.

Finally, with respect to *ellipsis* and *terminology*, we found them to be the least tagged issues (Table 6), not showing as the most tagged ones in any domain (Table 7). However, while ellipsis has most of the context span being preceding sentences - which is not so surprising given that ellipsis is a form of anaphora, and therefore, has its first mention generally before the ellipsis - terminology, interestingly, had most of the its context span being tagged as global, where the reader needs the whole boundary of the text to be able to disambiguate it. It is interesting to notice that in the news domain, ellipsis was the second issue to require the longest context span (5.28 sentences long on average with a median of 2 sentences, Table 4).

Regarding the domains, results indicate that the context span necessary to solve these context-related issues highly depend on the domains as it is the case for literature and subtitles which have presented the longest context spans. We note that this does not seem to be related to the length of the sentences in the corpus, since the average sentence length for the literature domain (Table 3) is the shortest in the corpus.

We believe that our findings will help the NLP and MT communities who are seeking to solve the issues of adding more context to their system, and shed some light on how these issues behave in different domains, regarding the length of the context necessary as well as the position they can be found. Moreover, we show that the standard approach of relying on only two preceding sentences as context might not be enough depending on the domain and issue types.

6. Acknowledgements

This project was funded by the Irish Research Council (GOIPD/2020/69). ADAPT, the Science Foundation Ireland Research Centre for AI-Driven Digital Content Technology at Dublin City University, is funded by the Science Foundation Ireland through the SFI Research Centres Programme (Grant 13/RC/2106_P2).

7. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR*, San Diego, CA.

Baker, M. (2006). Contextualization in translator-and interpreter-mediated events. *Journal of pragmatics*, 38(3):321–337.

Castilho, S., Popović, M., and Way, A. (2020). On Context Span Needed for Machine Translation Evaluation. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France, may.

Castilho, S. (2020). On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1150–1159, Online, November. Association for Computational Linguistics.

Castilho, S. (2021). Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 34–45, Online, April. Association for Computational Linguistics.

Grosjean, F., Dommergues, J.-Y., Cornu, E., Guillelmon, D., and Besson, C. (1994). The gender-marking effect in spoken word recognition. *Perception & Psychophysics*, 56(5):590–598.

Lopes, A. V., Amin Farajian, M., Bawden, R., Zhang, M., and Martins, A. F. T. (2020). Document-level

Neural MT: A Systematic Comparison. In *22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal.

Melby, A. and Foster, C. (2010). Context in translation: Definition, access and teamwork. *The International Journal for Translation & Interpreting Research*, 2, 11.

Rikters, M. and Nakazawa, T. (2021). Revisiting context choices for context-aware machine translation.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Proceedings of NIPS*, pages 3104–3112, Montreal, Canada.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 5998–6008, Long Beach, CA, December.

Wang, L. (2019). *Discourse-aware neural machine translation*. Ph.D. thesis, Ph. D. thesis, Dublin City University, Dublin, Ireland.

Yin, K., Fernandes, P., Pruthi, D., Chaudhary, A., Martins, A. F. T., and Neubig, G. (2021). Do context-aware translation models pay the right attention? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 788–801, Online, August. Association for Computational Linguistics.

8. Language Resource References

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana, June. Association for Computational Linguistics.

Cai, X. and Xiong, D. (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, Suzhou, China, December. Association for Computational Linguistics.

Castilho, S., Cavalheiro Camargo, J. L., Menezes, M., and Way, A. (2021). Dela corpus-a document-level corpus annotated with context-related issues. In *Proceedings of the Sixth Conference on Machine Translation*, pages 571–582. Association for Computational Linguistics (ACL), November.

Guillou, L., Hardmeier, C., Lapshinova-Koltunski, E., and Loáiciga, S. (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, Belgium, Brussels, October. Association for Computational Linguistics.

Müller, M., Rios, A., Voita, E., and Sennrich, R. (2018). A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium, October. Association for Computational Linguistics.

Rysová, K., Rysová, M., Musil, T., Poláková, L., and Bojar, O. (2019). A test suite and manual evaluation of document-level NMT at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 455–463, Florence, Italy, August. Association for Computational Linguistics.

Voita, E., Sennrich, R., and Titov, I. (2019). 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, July. Association for Computational Linguistics.

Vojtěchová, T., Novák, M., Klouček, M., and Bojar, O. (2019). SAO WMT19 test suite: Machine translation of audit reports. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 481–493, Florence, Italy, August. Association for Computational Linguistics.