Engaging Content
Engaging People

A World Leading SFI Research Centre

Science Foundation Ireland For what's next

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

Sheila Castilho
Dublin City University

Why do we need context?

- Document-level evaluation of MT allows for a more thorough examination of the output quality with context.
- Lack of definition of what constitutes ‘document-level’ with different researchers using different spans.

The DELA Corpus (Castilho et al 2021)

Six domains (literary, review, news, subtitles, medical, legislation) annotated with issues that occur in a translation from EN into PT-BR when no context information is given:

- **Reference:** 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:** 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:** a gender ambiguity was unsolvable within the sentence itself, e.g.: *I'm surprised to see you back so early*. Gender →surprised = feminine
- **Number:** a number ambiguity within the referential chain, e.g.: *I was praying for you*. Number → you = plural
- **Lexical Ambiguity**: 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:** a wrongly domain-specific word translation caused by contextual poor sentences, 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 (par

Types of Context

The *shortest context* span necessary to solve every issue annotated have been 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.
- **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, 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.

Context Position

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	-	

- Majority PREC, FOLL
- Diverse context span for lexical ambiguity
- Most common: gender, lexical ambiguity

Context Span

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

- Average PREC tends to be longer than the average FOLL context span
- Longest average span : gender and lexical ambiguity

Summary per Domain

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

Gender: one of the most tagged issues in every domain

Lex. Amb: the most tagged in the literary and medical domains, being also one of the three most tagged issues in the review and news.

Reference: third most annotated issues in the corpus, one of the shortest average context length (reviews and news domain)

Number: most tagged in the subtitle domain, with the longest context span needed

Ellipsis & Terminology: least tagged ones. Ellipsis the second longest context span in the news domain.

Conclusion

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.

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 is the shortest in the corpus

- 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.

DCU Ollscoil Chathair Bhaile Átha Cliath Dublin City University

FUNDED BY:
Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union European Regional Development Fund
Science Foundation Ireland For what's next

#LoveIrish Research
IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn