Probing Coreference Models of Romance Languages using Multiword Expressions

Evelin Amorim
INESC TEC / Porto, Portugal
evelin.f.amorim@inesctec.pt

Relevant UniDive working groups: WG3

1 Introduction

The nominal coreference resolution task is dedicated to identifying *who* is an entity mentioned by some expression in a text (Jurafsky and Martin, 2000). For instance, consider the sentence: "John saw Mary in the market, and she was wearing a red dress". In this example, "she" refers to the entity "Mary". This is a straightforward nominal coreference relation between a pronoun and an entity. However, some languages drop the subject in some sentence constructions and are known as pro-drop languages (Saab, 2016). Among these languages, there are the romance languages, which encompass Spanish, Portuguese, and Italian. We should highlight that although French is a romance language, it is not pro-drop. Nonetheless, we intend to consider this language in our planning.

The dropping of pronouns in those languages can arise in coreference[1] situations, which can influence the predictions of a zero-show or a few-show coreference model. Consider the following example of coreference in Portuguese that comprises a case where a definitive article (os/the) refers to an entity (Os usuários antigos/Old Users)[2].

Example 1.1 (Portuguese Coreference Example) *Os usuários antigos eram os que mais reclamavam.*

Old users were the ones who complained the most.

This is a usual construction in the Portuguese language where no pronoun or entity refers to an entity. If we look at the English version of the sentence, there is a complete subject, instead of only an article, referring to the entity "Old users", which is "the one". This is particularly common in romance languages, except for the French language.

In addition to this challenge in the coreference task, there is the possible influence of Multiword Expression (MWEs). These linguistic expressions are specific and idiosyncratic for each language (Savary et al., 2018) and can be categorized according to their ability to be coreferential (Éric Laporte, 2018). Considering this relevance for coreference, we aim to investigate the influence of MWEs in coreference models in romance languages.

MWE and pro-drop features in some of the romance languages pose some challenges to the coreference task, hence we propose the following research questions:

(Q1): Are the current state-of-the-art coreference models for the English language, when applied as zero-shot inference models in Romance languages, consistently producing similar errors across these languages?

(Q2): Do few-shot coreference models trained in Romance languages and tested in the same languages will exhibit similar errors?

(Q3): If errors produced by zero-shot and few-shot coreference models are consistent across Romance languages, can these errors be categorized, and can they provide insights for enhancing few-shot or even zero-shot coreference models for the Romance languages?

(Q4): To what extent can Multi-Word Expressions (MWEs) influence the prediction errors of zero-shot and few-shot coreference models, and can this influence be quantified in Romance languages?

2 Related Work

Savary et al. (2023b) conducted a study that assessed the occurrences of MWE in coreference chains for the French language. The investigation is the most similar research to our proposal

[1] From this point on in the paper, we will always consider the coreference task as the nominal coreference task.

[2] A glossing for this example is: Os usuários antigos/The old users, eram/were, os/the, que/who, mais/the most, reclamavam/complained.

since it analyzes the influence of MWE in a coreference dataset in the French language. The results of the study show that MWEs do not occur in long-chain coreferences. A more linguistic and qualitative analysis of MWE and coreference was performed by Éric Laporte (2018). The author proposed a taxonomy for MWE based on coreference, phrase structure, and other linguistic properties. Some examples supported the proposed taxonomy. Other works, however, apply a coreferential model to the SemEval-2010 Task 1 dataset (Recasens et al., 2010), which comprises three romance languages: Spanish, Catalan, and Italian. For instance, (Bohnet et al., 2023) and (Le and Ritter, 2023). However, none of them scrutinize the state-of-the-art coreference models regarding the errors to achieve better results. In this plan, we propose to do this investigation, in addition to researching the influence of MWE in the predictions. This can be shedding new light on the influence of MWE and how LLMs behave in the Romance languages.

3 Planned Methodology

The planned methodology for this research is guided by the research questions. Following, we detail the list of the programmed tasks for each research question.

Q1. For this research question, we have the following list of tasks. **(1)** *Reviewing of the state-of-the-art for the coreference task.* This is an already done task. Our reference works are Liu et al. (2022); Bohnet et al. (2023); Zhang et al. (2023), which all modeled the coreference task as a sequence-to-sequence model. However, we intend always to keep up with the current news in the field. **(2)** *Collecting of coreference datasets for Romance languages.* The start point for this task is the following works for the Portuguese language (Lapshinova-Koltunski et al., 2022; Vieira et al., 2018), the Spanish language (Recasens and Martí, 2010), the Italian language (Rodrıguez et al., 2010; Guarasci et al., 2021), and the French language (Wilkens et al., 2020). **(3)** *Applying of pre-trained English coreference model to the collected datasets.* The initial experiments we are planning to do are zero-shot, i.e., no training in the romance languages. The model we intend to apply is the model proposed by Bohnet et al. (2023) since is the state-of-art for the English coreference and the source code for training is available. **(4)***Qualitative Analysis of the errors of the prediction.* We intend to analyze a sample of randomly selected errors produced by the model in the previous task.

Q2. The task for this research question leverages the outcome of the tasks assigned to the previous research question, although we intend to do a deeper analysis of the errors and also report them.

Q3. Like the previous questions, we intend to get the outcome of the tasks listed for research question one. However, here we intended to add the following tasks to the planned work. **(1)** *Reviewing about the taxonomy of coreference errors.* Is there in the literature a proposed taxonomy of errors or can you employ the conference taxonomy to categorize the errors? **(2)***Analysis of error categories for possible improvements.* Given the error categories, are there any improvements that we can make to the current models for romance languages?

Q4. For this research question, we planned the following tasks. **(1)** *Collecting of lexical resource databases of MWEs for Romance languages.* We will begin this task using the mwetoolkit (Ramisch et al., 2010), which is a toolkit to aid the extraction of MWEs. Other possible references are (Madabushi et al., 2022; Savary et al., 2023a). **(2)** *Extraction of MWEs in the coreference corpora.*Next, we going to map the occurrences of MWEs in the coreference corpora. To do this automatically, we can detect MWEs in the coreference corpora by (a) a lexical matching methodology or by (b) modeling the task as a sentence classification task, like in the Shared Task from SemEval 2022 Task 2 (Madabushi et al., 2022). We aim to test both methodologies. After this task, we will be able to assess which languages we could probe the coreference models. **(3)** *Analyze the coreference by MWEs.* Using the results of the previous task, and the results from task 3 of question **Q1** (results from the application of the coreference English model), we going to analyze the influence of MWEs in the prediction of coreference models in the Romance languages. **(4)** *Development of new models or strategies.* After the analysis of the previous task, we aim to develop new strategies to improve the coreference models for the romance languages. The results and the models of this step will be released publicly.

4 Expected Results

For this research plan, the expected results are (1) shedding light on the current state-of-art corefer-

ence models for romance languages; (2) mapping the influence of MWE in the coreference models for romance languages, and (3) improvements in the coreference models for romance languages.

References

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Raffaele Guarasci, Aniello Minutolo, Emanuele Damiano, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2021. Electra for neural coreference resolution in italian. *IEEE Access*, 9:115643–115654.

Daniel Jurafsky and James H Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.

Ekaterina Lapshinova-Koltunski, Pedro Augusto Ferreira, Elina Lartaud, and Christian Hardmeier. 2022. Parcorfull2. 0: A parallel corpus annotated with full coreference. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 805–813.

Nghia T Le and Alan Ritter. 2023. Are large language models robust zero-shot coreference resolvers? *arXiv preprint arXiv:2305.14489*.

Tianyu Liu, Yuchen Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structured prediction with language models.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. Semeval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. *arXiv preprint arXiv:2204.10050*.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: A framework for multiword expression identification. In *LREC*, volume 10, pages 662–669. Valletta.

Marta Recasens, Lluís Màrquez, Emili Sapena, M Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 1–8.

Marta Recasens and M Antònia Martí. 2010. Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Language resources and evaluation*, 44:315–345.

Kepa Joseba Rodrıguez, Francesca Delogu, Yannick Versley, Egon W Stemle, and Massimo Poesio. 2010. Anaphoric annotation of wikipedia and blogs in the live memories corpus. In *Proceedings of LREC*, pages 157–163.

Andrés Saab. 2016. On the notion of partial (non-) pro-drop in romance. *The morphosyntax of Portuguese and Spanish in Latin America*, pages 49–77.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023a. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Cebiroğlu Eryiğit, Voula Giouli, Maarten van Gompel, et al. 2018. Parseme multilingual corpus of verbal multiword expressions. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*.

Agata Savary, Jianying Liu, Anaëlle Pierredon, Jean-Yves Antoine, and Loïc Grobol. 2023b. We thought the eyes of coreference were shut to multiword expressions and they mostly are. *Journal of Language Modelling*, 11(1):147–187.

Renata Vieira, Amália Mendes, Paulo Quaresma, Evandro Fonseca, Sandra Collovini, and Sandra Antunes. 2018. Corref-pt: A semi-automatic annotated portuguese coreference corpus. *Computación y Sistemas*, 22(4):1259–1267.

Rodrigo Wilkens, Bruno Oberle, Frédéric Landragin, and Amalia Todirascu. 2020. French coreference for spoken and written language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 80–89.

Wenzheng Zhang, Sam Wiseman, and Karl Stratos. 2023. Seq2seq is all you need for coreference resolution.

Éric Laporte. 2018. Choosing features for classifying multiword expressions. In Manfred Sailer and Stella Markantonatou, editors, *Multiword Expressions: Insights from a Multi-lingual Perspective*, pages 143–186. Language Science Press, Berlin.