Nominal Coreference Annotation in IberEval2017: the case of FORMAS group

Marlo Souza[1], Rafael Glauber[1], Leandro Souza de Oliveira[1], Cleiton Fernando Lima Sena[1], and Daniela Barreiro Claro[1]

Institute of Mathematics and Statistics, Federal University of Bahia - UFBA,
Av. Adhemar de Barros, S/N, Ondina - Salvador-BA, Brazil,
msouza1@ufba.br, rglauber@dcc.ufba.br, leo.053993@gmail.com,
cflsena2@gmail.com, dclaro@ufba.br

Abstract. This work describes the participation of the FORMAS group from Federal University of Bahia (UFBA) in the Shared Task on Collective Elaboration of a Coreference Annotated Corpus for Portuguese Texts for IberEval 2017. As such, it describes the creation of a corpus annotated with coreference information for the Portuguese language. We discuss the choices adopted oin the annotation process, as well as the results obtained and their possible application to the development of methods and systems focusing on the processing of texts in portuguese.

1 Introduction

Anaphora and coreference resolution are well-established problems in the literature of Computational Linguistics [17, 11, 9, 14]. While there is some terminological confusion regarding the use of anaphora resolution and coreference identification, in this work we adopt these terms to be similar and to refer to the problem commonly known as anaphora resolution in the computational linguistics literature. We consider the problem of nominal coreference resolution as the problem concerning the identification of two (or more) nominal phrases which refer to the same discourse entity in the domain of discourse [13].

This work describes the creation of a corpus annotated with coreference information in the context of the Shared Task on Collective Elaboration of a Coreference Annotated Corpus for Portuguese Texts for IberEval 2017. Our team was composed of five researchers, with three main annotators.

2 The problem of identifying coreference chains

It has been pointed out in the literature that the problem of coreference resolution and the guidelines for corpus annotation for this task are underspecified, or that, at least, there are some terminological inadequacies in their definition [18, 13]. Thus, as a first step in our group's annotation effort, we tried to establish a common understanding of the phenomenon and the difficulties regarding the annotation process.

Commonly, two noun phrases (NPs) are said to co-refer if they "refer to the same entity" [8]. This definition requires of coreferring noun phrases the properties that (i) they refer directly to an (unique, unambiguous) entity and (ii) this entity is identifiable from the context of the noun phrases. These requirements, however, are true only for a small set of noun phrases in a text.

On the notion of referring used in this work, while the semantic/philosophical logic notion of referring is a relation between a linguistic expression and an object, it is clear that this relation is usually too restrictive to explain the notion we are interested in this work. Otherwise, NPs contained in counterfactual or hypothetical statements, as (1) below, would be non-referring, even if the noun phrases 'a car' and 'it' are naming the same entity in the universe of the discourse, i.e. a hypothetical car. As such, in this work we adopt a broader notion of referring, which holds between two linguistic expressions. In this case it is not problematic to say that the pronoun 'it' in sentence (1) refers to the same entity as the NP 'a car'. The we adopt the definition that two NPs co-refer if they refer to the same entity introduced in the universe of discourse, i.e. have the same discourse referent in the nomenclature of functional grammar theory [6].

(1) If I had **a car**, I would drive it to the coast.

Notice that noun phrases may have several semantic functions in a sentence, according to Poesio [13], and not all noun phrases refer directly to an entity. To understand which kind of NP may be of importance to the annotation task, we must investigate further the uses of NPs in the language. Some of the functions a NP may have in a sentence are:

- Referring: a NP is said to be referring if it refers directly to some discourse entity, as the NP "a car" in sentence (1).
- Quantificational: a noun phrase may acts as a quantification over the domain of discourse bounding the interpretation of the predicate to a set of discourse entities denoted by the NP. For example, consider the NP "Every TV network" in sentence (2) below. This NP act as a quantification ranging over all discourse entities which are considered to be TV networks. As such, the (logical) meaning of the sentence (2) may be expressed by the logical expression (2').

(2) **Every TV network** reported its profits.
(2') $\forall x.(Tv\_network(x) \rightarrow \exists y.(reported(x, y) \wedge profits\_of(y, x)))$

Notice that, as Van Deemter and Kibble [18] point out, we cannot simply take the quantificational NP "Every TV network" in sentence (2) to directly refer to the class of all TV network entities in the universe of discourse, otherwise the (anaphoric) relation between the reference of this NP and the pronoun 'its' cannot be properly established. For instance, if we take 'its' to co-refer with 'Every TV network', the sentence (2") below would be a paraphrase of (2).

(2") Every TV network reported every TV network's profits.

– Predicative: a NP can express properties of an object, and it can not refer to a specific entity as in the case of 'a preacher' in sentence (3) provided by Poesio [13], which describes a property (i.e. a predicate) of the entity referred by 'Kim'. As such, the (logical) meaning of sentence (3) can be expressed by the logical expression (3').

(3) Kim is a preacher.
(3') $preacher(Kim)$

– Expletive: some languages require the presence of certain verbal arguments, such as French or English in which null subjects are not allowed in certain types of sentences. In these languages, non-referring NPs may be used as fillers, to occupy a required syntactical position in the sentence. This is the case of the pronoun 'It' in sentence (4) for the English language and similarly the pronoun 'Il' in sentence (5) for the French language.

(4) It's two o'clock.
(5) Il est deux heures.

From this discussion, it is clear that only those NPs that refer to some entities in the domain of discourse are of interest to the annotation process, namely those fuunctioning as referring NPs and as quantificational NPs. As Poesio [13] points out, however, it is not always clear how to classify a given noun phrase in a sentence according to their semantic function, even for a human [15]. More yet, usually there are different ways to interpret a given NP depending on the adopted linguistic theory.

Another aspect of identification of coreferent NPs concerns how to delimit when the referents of two NPs can be considered equal. Notice that the relation between the entities referred by two distinct NPs may not be that of identity, and yet it is arguable the case that the two NPs to corefer. Let's consider the case of the case of the NPs "The house" and "the bathroom" in the sentence (6) below.

(6) The house is great, but the bathroom is too dark and humid.

The referents of these two NPs are related by a meronymy relation, i.e. the bathroom to which the second NP refers is a part of the house to which the first refers. As such, the NPs refers to the same entity, but to different parts of it. This kind of coreference, which were subject to anotation in the MUC-7 task [7], is often called associative coreference.

In the annotation task described in this work, we do not consider associative coreferences. We aim to annotate only the cases of coreference established by means of referring NPs and by quantificational NPs. However, as Van Deemter and Kibble [18] point out, it is not always clear how to establish the reference of a quantificational NP. On one hand, if we establish that the referent of a quantificational NP, such as "Every TV network" in sentence (7) we lose the coreference relation between this NP and the pronoun "its" in the same sentence.

If we take the meaning to be a single TV Network bound to the context of the quantification, we may not establish the connection between "Every TV network" and the pronoun "they" in the sentence (7) below.

(7)**Every TV network** reported its profits. They are required by the government to do so.

This is not an easy problem to fix. Particularly, depending on the context, either option in defining the referent of the NP may be more suitable. Since our aim in this annotation is to maximize the annotation of coreferent NPs, we establish that either possibility may be taken by the annotator, as long as it is done consistently throughout the text. In other words, the annotator may choose either that the referent of "Every TV network" is the same as "it" or the same as "they", as long as the annotator does not change the referent while analyzing the text.

Regarding possible difficulties relating change over time, for descriptors like "the president of Brasil" for which the reference is dependent on a temporal context, we adopted the same strategy to that of the MUC-7 conference, i.e. "two markables should be recorded as coreferential if the text asserts them to be coreferential at ANY TIME" [7, p. 11], as long as it is clear by context that the reference of the NPs is intended to be the same.

3 The corpus

To perform the annotation task, we composed a corpus of thirty encyclopedic texts written in portuguese language, taken from the Wikipedia [1]. Wikipedia texts are an important resource for languages with scarce computational linguistic resources, since they compose a corpus of a significant size, usually coupled with important annotation, such as the domain classification, cross-references between pages in different languages, etc.

Wikipedia corpora have been widely used in the NLP literature for its availability, structure and existing metadata. For the portuguese language, particularly, the Wikipedia Corpus has been used in Ontology Learning [20], Open Information Extraction [21, 2, 22, 12], Named Entity Recognition [3, 19], as well as been subject of the Págico - Shared Task on information retrieval [16], among many others.

The texts that compose the corpus used in the annotation process were randomly selected from the Wikipedia dump of March 26 of 2017 (03.26.2017) using the Wikipedia Extractor tool[2]. To select the texts we have established the following criteria:

1. the text must have approximatively 1200 (between 1100 and 1400) words;
2. the text must not concern physical or mathematical theories, nor contain mathematical formulas as figures;

[1] http://pt.wikipedia.org
[2] Available at: `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor`

3. the topic of the text must not be wiki meta-information, such as discussion pages;
4. the text must be a running text discussion of a topic, excluding thus any Wikipedia page containing lists (e.g. List of awards received by Justin Timberlake).

The first requirement was made to conform to the shared task specification of texts containing 1200 words. The second requirement is justified by the fact that complex mathematical formulas are not easy to parse by text processors and, in fact, are excluded from the text in the extraction using the Wikipedia Extractor. Since the pages describing mathematical and physical theories commonly rely in a great amount of mathematical formulas to describe their topics, the extracted text becomes poorly informative and difficult to process. The third and fourth requirements are made to exclude uninteresting pages which are become common in the corpus considering the restriction of texts with size of 1200 words.

4 The annotation tool

Following the methodology established for the shared task, the annotation process consisted of two steps. In the first step, an initial automatic annotation of the corpus was performed by the organizing team, in which each NP in a text is identified and those NPs participating in a coreference chain are grouped together. In the second step, each annotation team performed the manual correction of the initial annotation, both of the problems of delimitation of NPs and the identification of coreference chains. To perform this manual correction, the annotation teams used the tool CorrefVisual [5], provided by the organizing team. Regarding our experience with the tool and the annotation process, we offer some considerations for the task and the resulting corpus.

First, it was our impression, in the annotation process, that the visual grouping of the elements in the same coreference chain, as provided by the annotation tool, does indeed help the verification that all noun phrases in the same chain are actually coreferent. When the number of coreference chains grew, however, the navigation through all these groupings became a hindrance in the annotation. Since the tool only provided manual navigation or text search, deciding whether a given NP should be included in some existing chain usually involved navigating through several coreference chains, which became very time-consuming.

Regarding the noun phrase identification and delimitation, while the tool allowed the correction of the boundaries of noun phrases, in some cases the noun phrases were not even partially identified by the tool. Since the annotation tool did not allow the creation of new noun phrases, only to alter the boundaries of those already identified, some coreference relations were not possible to be annotated. This was a common occurrence in the presence of compound noun phrases such as “os gêneros Sambucus e Viburnum” (the genera Sambucus and Virbunum), as the tool normally presented only either the option with the complete NP “os gêneros Sambucus e Viburnum” or two separated noun phrases “Sambucus” and “Viburnum”.

While it is our belief that the annotation tool did indeed help the annotation process, reducing the amount of labor involved in it, we also believe that the amount of restrictions imposed by the tool to the annotators may have an impact in the quality of the resulting corpus.

5 Annotation evaluation

The agreement in the annotation was measured by means of the Kappa statistics. Four texts were annotated by all three annotators for this comparison. The agreement for each text and among the group is depicted in Table 1.

Table 1. Agreement by the annotators measured by Kappa statistics

Text	**Agreement**
text 1	0.78
text 2	0.23
text 3	0.40
text 4	0.35
Group Kappa	0.43

To understand these results, we performed a quantitative analysis of the texts to determine the reason for the high deviation in the agreement among the texts. The hypothesis was that the higher agreement was achieved in simpler texts, while the lowest agreements were achieved on more complex ones. For this quantitative analysis, we evaluated the number of noun phrases in the text (**#NPs**), as well as the statistics of the annotation, such as number of identified chains (**#chains**), number of NPs that had a coreference relationship with another (**#correferent**), the average size of the coreference chains in the text (**Avg size of chains**) and the size of the biggest identified chain (**Size of biggest chain**) based on the annotation of each annotator of the group. The resulting data are depicted in Tables 2, 3 and 4

Table 2. Statistics of the texts and annotation according to annotator 1

Text	**#NPs**	**#chains**	**#correferents**	**Avg size of chains**	**Size of biggest chain**
text 1	422	25	106	4.24	35
text 2	452	45	175	3.88	19
text 3	407	42	191	4.55	27
text 4	479	61	199	3.26	30

Notice that in text 1, the text with higher agreement, the amount of identified coreference chains in small (for all annotators) compared to the others and the

Table 3. Statistics of the texts and annotation according to annotator 2

Text	#NPs	#chains	#correferents	Avg size of chains	Size of biggest chain
text 1	422	17	85	5.0	32
text 2	452	29	129	4.45	45
text 3	407	28	130	4.64	22
text 4	479	75	295	3,93	30

Table 4. Statistics of the texts and annotation according to annotator 3

Text	#NPs	#chains	#correferents	Avg size of chains	Size of biggest chain
text 1	422	7	48	6.85	29
text 2	452	11	46	4.18	13
text 3	407	7	43	6.14	13
text 4	479	29	87	3	18

number of NPs that participate in a coreference chain is also slightly smaller than for the other texts, which seems to agree with our hypothesis.

While the increase in the number of identified chains in text 4 did reduce the agreement, this behavior was not observed in text 2, which has the lowest agreement and almost the same number of identified chains as text 3. Also, it is not clear that the increase from 25 coreference chains with 106 coreferent NPs in text 1 to 42 chains with 191 coreferent NPs in text 3 (for annotator 1) could explain such a significant variation in agreement between the two texts.

Analyzing text 3, we identified that this text suffers from extreme poor writing quality, being, in fact, a translation from the English language with several only partially translated expressions within it. In this context, both the NP delimitation as well as the coreference determination became compromised, which explain the low value in agreement for the annotators. As such, we consider that the text should be treated as an outlier and non representative of the result of the annotation.

Regarding the general results for each annotator, notice that annotators 1 and 2 are more coherent with each other in their annotations, identifying similar number of coreference chains and coreferent NPs for all texts, while annotator 3 deviates more from the other two. One possible explanation for such behavior is that, apart from the inherent difficulty of establishing reference of NPs, the notion of coreference adopted in the work may not have been a consensus for all annotators.

6 Coreference information in Open Information Extraction

Open Information Extraction (Open IE) is the area that studies methods for extracting information from fragment texts without any previous constraint on the kind of relations to be identified. It was introduced by Banko et al. [1] with

the system TextRunner and has flourished into an active area of research in Natural Language Processing.

Early Open IE method relied on using domain-independent extraction patterns to identify relation instantiations [10]. As a result, several extractions made by this systems have low quality, the result of what Etzioni et al.[4] call incoherent and uninformative extractions. According to Etzioni et al.[4], incoherent extractions are those which have no meaningful interpretation, while uninformative extractions are those in which critical information for the interpretation of the expression is missing. These authors claim that methods that tackle the problem of reducing these low quality extractions compose the second generation of Open IE systems.

Notice that resolving coreference is an essential challenge to guarantee the minimization of uninformative extractions of Open IE systems. The reason for that is that outside its discursive context, pronouns and other descriptors of discourse entities have no clear referent. Let's analyze the text (6) below.

(6) Mariana's car is more reliable than that of Louis. She takes very good care of it. The car was revised this week

Typical examples of extracted information from Open IE systems, considering the text of (6) would be the tuples represented in (6*),(6**) and (6***).

(6*) (Mariana's car, is more reliable, that of Louis)
(6**)(She, takes good care of, it)
(6***) (The car, was revised, this week)

Without its linguistic context, the extractions (6**) and (6***) are uninformative, since it is not clear to which entities the pronouns "She" and ''it" refer to in sentence (6**), nor to which car (Mariana's or Louis') the NP "The car" refers to in sentence (6***).

With the construction of a corpus of coreference, we aim to allow the development of Open IE systems for the Portuguese language which explore coreference information to extract more informative relations without while obtaining the information that would be lost if we discarded the extractions (6**) and (6***).

7 Conclusion

The present work described the participation of the FORMAS group in the shared task for the collective elaboration of a coreference annotated corpus for Portuguese texts at IberEval 2017. In this work, we discussed the notion of coreference adopted by our group for the annotation process, the corpus we used as well as our experience during the annotation process.

We believe that, while the annotation showed moderate agreement from the annotators, the resulting corpus can (and will) be an important resource to the Portuguese language, allowing the development of interesting methods and systems focusing on this language, considering the inherent difficulty of solving this

problem and the scarcity of resources for the Portuguese language. Particularly, we plan to evaluate the resulting corpus within the context of Open IE in the near future, to measure how such a resource can foster the development of methods that minimizes the extraction of uninformative relations, while being able to extract all the information from a text.

In regard to the shared task, we consider that the notion of coreference adopted expected in the shared task was not clearly defined, and, as such, we felt the necessity to explicit the notions and choices adopted by our groups. As a result of the underspecification of the task, it is not clear to us how the many corpora generated by different groups participating in the shared task can be united to form a corpus for coreference resolution in the Portuguese language. We believe that, to merge all these corpora, it will be necessary to identify possible inconsistencies in the annotation processes, apart from the pure quantitative evaluation of inter-annotator agreement.

References

1. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: IJCAI. vol. 7, pp. 2670–2676 (2007)
2. Batista, D.S., Forte, D., Silva, R., Martins, B., Silva, M.: Extracçao de relaçoes semânticas de textos em português explorando a dbpédia e a wikipédia. linguamatica 5(1), 41–57 (2013)
3. Cardoso, N.: Rembrandt-reconhecimento de entidades mencionadas baseado em relaçoes e análise detalhada do texto. In: Encontro do Segundo HAREM (2008)
4. Etzioni, O., Fader, A., Christensen, J., Soderland, S., Mausam, M.: Open information extraction: The second generation. In: IJCAI. vol. 11, pp. 3–10 (2011)
5. Fonseca, E., Sesti, V., Vanin, A., Vieira, R.: Guia para anotação de correferência. http://ontolp.inf.pucrs.br/coref/ibereval2017/CorrefVisual.zip (2017)
6. Gundel, J.K., Hedberg, N., Zacharski, R.: Cognitive status and the form of referring expressions in discourse. Language pp. 274–307 (1993)
7. Hirschman, L., Chinchor, N.: Muc-7 coreference task definition. In: MUC-7 Proceedings. Science Applications International Corporation
8. Hirschman, L., Robinson, P., Burger, J., Vilain, M.: Automating coreference: The role of annotated training data. In: Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing. pp. 118–121 (1997)
9. Hirst, G.: Discourse-oriented anaphora resolution in natural language understanding: A review. Computational Linguistics 7(2), 85–98 (1981)
10. Li, H., Bollegala, D., Matsuo, Y., Ishizuka, M.: Using graph based method to improve bootstrapping relation extraction. Computational linguistics and intelligent text processing pp. 127–138 (2011)
11. Mitkov, R.: Anaphora Resolution. Longman (2002)
12. Pires, J.C.B.: Extração e Mineração de Informação Independente de Domínios da Web na Língua Portuguesa. Master's thesis, Universidade Federal de Goiás (2015)
13. Poesio, M.: Linguistic and cognitive evidence about anaphora. In: Anaphora Resolution, pp. 23–54. Springer Berlin Heidelberg (2016)
14. Poesio, M., Stuckardt, R., Versley, Y.: Anaphora resolution (2016)
15. Poesio, M., Vieira, R.: A corpus-based investigation of definite description use. Computational linguistics 24(2), 183–216 (1998)

16. Santos, D.: Porquê o págico? razoes para uma avaliaçao conjunta. Linguamática 4(1), 1–8 (2012)
17. Van Deemter, K., Kibble, R.: On coreferring: Coreference in muc and related annotation schemes. Computational linguistics 26(4), 629–637 (2000)
18. Van Deemter, K., Kibble, R.: On coreferring: Coreference in muc and related annotation schemes. Computational linguistics 26(4), 629–637 (2000)
19. Weber, C.: Construção de um corpus anotado para classificação de entidades nomeadas utilizando a Wikipedia e a DBpedia. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul (2015)
20. Xavier, C.C., De Lima, V.L.S.: A semi-automatic method for domain ontology extraction from portuguese language wikipedia's categories. In: Brazilian Symposium on Artificial Intelligence. pp. 11–20. Springer, Berlin, Heidelberg (2010)
21. Xavier, C.C., de Lima, V.L.S., Souza, M.: Open information extraction based on lexical-syntactic patterns. In: Intelligent Systems (BRACIS), 2013 Brazilian Conference on. pp. 189–194. IEEE (2013)
22. Xavier, C.C., de Lima, V.L.S., Souza, M.: Open information extraction based on lexical semantics. Journal of the Brazilian Computer Society 21(1), 4 (2015)