Adapting an Entity Centric Model for Portuguese Coreference Resolution

Evandro B. Fonseca[1], **Renata Vieira**[1], **Aline Vanin**[2]
[1]PUCRS University – Porto Alegre – Brazil
[2]UFCSPA University – Porto Alegre – Brazil
evandro.fonseca@acad.pucrs.br, renata.vieira@pucrs.br, aline.vanin@ymail.com

Abstract

This paper presents the adaptation of an Entity Centric Model for Portuguese coreference resolution, considering 10 named entity categories. The model was evaluated on named e using the HAREM Portuguese corpus and the results are 81.0% of precision and 58.3% of recall overall, the resulting system is freely available.

Keywords: Coreference Resolution, Information Extraction, Rule-based Models

1. Introduction

Coreference resolution is a well known challenge in computational linguistics. While we could imagine that it is relatively easy to identify automatically that there is a link between identical or similar referents, such as in (1) "Barack Obama" and (2) "Obama", this might not be always the case. See (3) "Adalberto Portugal" and (4) "Portugal", for instance, whereas "Adalberto Portugal" refers to a person, "Portugal" might refer either to "Adalberto" or to Portugal, the country.

When dealing with Portuguese, this task is even more challenging, since resources are limited. Whereas Ontonotes (Pradhan et al., 2011), a coreference annotated corpus for English, Chinese and Arabic has around 34.290 coreference chains, a corpus with similar purposes for Portuguese, Harem (Freitas et al., 2010), has approximately 887 coreference chains.

In order to make available more resources for Portuguese, this paper presents the implementation and evaluation of an entity centric model for Portuguese, using a set of adapted rules, inspired by the Stanford Deterministic Coreference Resolution System (Lee et al., 2013). A rule based system was considered the best option, due to the shortage of annotated examples, necessary for learning approaches. We used the the Harem corpus (Freitas et al., 2010) to evaluate the adapted model. Our implementation is open source and we rely also on other open source resources, like Cogroo API (Silva, 2013).

The rest of this paper is organized as follows: Section 2 presents the related work; Section 3 describes our version of the system for Portuguese; in Section 4 we present the evaluation of the system; in Section 5 conclusions and future work are presented.

2. Related Work

Coreference Resolution is an old topic in NLP, but it is still challenging. There are many studies involving machine learning approaches. However, we consider that developing a rule based system would be more useful for Portuguese, since there is lack of a rich corpora, such as those employed to generate English models (see (Fernandes et al., 2014), (Martschat and Strube, 2015), among others). Instead, we used an available Portuguese corpus for evaluating the adapted model. Since, we are mainly interested in rule-based approaches we focus here in describing the most influential work for our system.

(Lee et al., 2013) use a deterministic approach to coreference resolution that combines the global information and precise features previously identified by machine-learning models with the transparency and modularity of deterministic, rule-based systems. Their Entity-Centric Model architecture applies a set of 10 deterministic sieves, where each sieve or model builds on the previous model's cluster output. Their model is based in two stages: mention detection, followed by clustering rules. In order to increase the recall, (Lee et al., 2013) combine several variations of matching rules. In addition, they include some "precise constructs", which may introduce semantic knowledge through appositive rules. They evaluate each rule independently, showing that each module introduce new levels of precision and recall. Lee et al.'s system was the winner in CONLL 2011 (Pradhan et al., 2011), solving coreferences for English, and it reached a F-measure of 61.0% (MUC metric).

Another related work which is also relevant for our study, is Garcia et al.'s Entity-Centric model (Garcia and Gamallo, 2014a). Garcia et al.'s system, or LinkPeople, is a model for coreference resolution of person entities. The model combines the multi-pass architecture and a set of constraints and rules. (Garcia and Gamallo, 2014a) use Lee et al.'s matching rules, and, in addition, they use a set of specific rules to deal with pronouns and person entities, as well as with linguistic reference phenomena, anaphora and cataphora. This system solves coreference (person entities) in three languages: Portuguese, Spanish and Galician, achieving 87.4% of F-measure for Portuguese, 91.7% for Galician and 88.8% for Spanish.

Most of the other related work for Portuguese dealing with nominal correference, proposes machine learning approaches, examples are (Coreixas, 2010), (Silva, 2011) and (Fonseca et al., 2015). Also, differently from most previous work for Portuguese, we make

available the system that is evaluated here.

3. Adapted Model

Our model is an adaptation of (Lee et al., 2013) for Portuguese (PT-BR). We adapted and implemented a set of sieves. The first two correspond to noun phrase extraction and pre-processing, a module that filters out some mentions, such as large NPs in (Lee et al., 2013). The other sieves are used to link two mentions if the conditions established by linguistic rules are satisfied. The sieves are described next:

1. NP_Extraction: The first module is responsible to extract noun phrases from a plain text. For this task we use Cogroo API((Silva, 2013)). The Cogroo API is a grammar checker for Portuguese that provides Part-of-speech, lemma and chunking annotation.

2. Pre-Processing: our pre-processing just removes numeric noun phrases ex: [200 farmers], [one million], [30^{o}F]...). We chose not to remove large mentions.

3. Exact_String_Matching: links the current NP to its antecedents when they are equal ex: [Carlos Nobre]. . . [Carlos Nobre].

4. Relaxed_String _Matching: This rule considers two mentions as coreferent if the strings previous to (and including) their heads are equal.

5. Appositive: if an NP is appositive with other - we consider appositives the noun phrases between comma, parenthesis and quotes.

6. Predicate nominative: this rule links two mentions when copulative subject-object relation exists. As in "[O carro] é [uma máquina incrível]" ([*the car*] is [*an awesome machine*]). In our adaptation, we consider only the verbs "ser" (*to be*) and "parece" (*to seem*).

7. Appositive Role: links two neighbour mentions if all the following constraints are satisfied: The current NP is a Proper name; the antecedent is a noun; the antecedent contains a determinant; the current NP does not contain a determinant. This rule helps to identify and link NPs like [[O telescópio] ([*the telescope*]) [Gemini]]. Different from (Lee et al., 2013), we use this rule for all NPs, not just person entities. In addition, we implemented a new clause, which process plural mentions: If the determinant is plural, all subsequent NPs that are proper names, separated by a comma or "e"(*and*), are linked with the previous noun.

   This links references such as [Os brasileiros] ([*The brazilians*]) with a list of names [Gilson Rambelli, Paulo Bava de Camargo e Flávio Rizzi], in a construction like " Os brasileiros, Gilson Rambelli, Paulo Camargo e Flávio Rizzi, pesquisadores..." (*the Brazilians..., researchers...*).

8. Relative Pronoun: links two adjacent mentions if the second NP is a relative pronoun. We use the following relative pronouns: "o qual", "cujo", "quanto", "quem", "que", "onde", considering variations in gender and number.

9. Acronym: This rule verifies if a mention is an acronym of the other. [A Organização das Nações Unidas] [ONU] ([*The United Nations*]). We used a simple strategy to build this rule. For all named entities we create two possible acronyms, considering all uppercase letters in the name. For example, for "Organização das Nações Unidas", we create "ONU" and "O.N.U.".

10. Strict_Head_Matching: this module consists in to link two mentions if any head word from the current mention matches with any head word from a previous mention. To avoid incorrect links some restrictions are applied, see Lee for details (Lee et al., 2013).

11. Variants_of_Strict_Head_Matching: This module applies different configurations of the Strict_Head_Match.

12. Proper_Head_Word_Matching: This module links two mentions if three conditions are satisfied: both noun phrases contain proper nouns; the proper nouns are equal; the mentions are not in an embedded construction (as in [Africa], [south of Africa]).

13. Relaxed_Head_Matching: This module relaxes the matching in Entity Head Match clause by allowing the mention head to match any word in the antecedent entity. This module is similar to Lee's, but in as in Exact_String_Matching and Relaxed_String Matching, uses the appositive role constraint. This way we avoid links between the mentions: [Miguel Guerra], [O agrônomo] ([*the agronomist* and [[*o* Advogado] ([*the lawyer*]) [Miguel]]. It is because in Relaxed_Head_Matching just one word from Head must match.

These were the adapted sieves. For the while, we chose to not implement Pronominal Coreference and Speaker Identification, since we are concerned with a more global coreference resolution model. Next we discuss the evaluation of the system.

4. Evaluation

The evaluation of the system is presented in three parts. The first is an automatic evaluation that relies

on Harem as the Gold Standard, an annotated corpus with named entities and their identity relations. Section 4.2 refers to the evaluation reported by other but similar approaches to correference resolution. Section 4.3 considers a manual analysis of the complete output of the system which includes other NPs, besides named entities.

4.1. Corpus based evaluation

An automatic evaluation was based on the Harem (Freitas et al., 2010) corpus, and reports results using the MUC metric (Vilain et al., 1995). HAREM (Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas) is an international shared evaluation for NLP systems for Portuguese. In its second edition, a task related to identity identification was proposed. Based on this task, HAREM provided a corpus with named entities and their identity relations annotation.The corpus (Freitas et al., 2010) has around 225k words. Relations between named entities were annotated considering 'identity' (our base for coreference), 'inclusion', and 'location' (occurs in). The annotation scheme is exemplified below:

< EM ID="ric-85133-257" CATEG="PESSOA" TIPO="INDIVIDUAL" > Italo Calvino < /EM >
< EM ID="ric-85133-290" CATEG="PESSOA" TIPO="INDIVIDUAL" COREL="ric-85133-257" TIPOREL="ident" > Calvino < /EM >

Each named entity has an entity id, a semantic category, 'Person', 'Location', 'Organization', among others (with subtypes); a relation descriptor, and coreference links between two or more entities. In the example we have a coreference link, between the entities "Italo Calvino" and "Calvino" ("ric-85133-257" and "ric-85133-290").
The corpus contains 7847 recognized named entities, distributed into 10 categories. These named entities represent a total of 887 correference chains, as presented in Table 1. Four our evaluation we used Harem instead of Garcia's corpus since the later is focused on person category, whereas in Harem we have a larger scope (10 named entity categories).
However, since Harem refers only to named entities, other noun phrases referring to the entities (such as: *the president*) are not considered in the corpus based evaluation. For these other cases we present a manual analysis in 4.2.
Table 1 shows the evaluation of the system for each named entity category. Our adapted model achieved F-measures above 70% for the majority classes: Person, Location and Organization. Time and Value presented the lowest results, since in pre-processing we discard all numeric noun phrases.For that reason, in the last line we show the results considering all named entity categories, except "Time" and "Value", in which there is an increase in recall and F-measure.
For the other classes we had F-measures above 60%. We consider that these are very promising results, considering other similar systems as discussed next.

Table 1: Experiment results by named entity categories.

	P	R	F	Chains
Person	80.7%	71.3%	75.7%	304
Loc	74.6%	82.8%	78.5%	179
Org.	66.8%	78.5%	72.1%	154
Work	78.7%	61.7%	69.2%	57
Event	57.5%	67.1%	61.9%	40
Thing	70.0%	75.4%	72.6%	48
Time	62.5%	37.5%	46.9%	42
Abstract	53.2%	76.2%	62.7%	40
Other	65.19%	61.4%	63.2%	11
Value	0.0%	0.0%	0.0%	12
All	81.0%	58.3%	67.8%	887
All-Time-Val	80.8%	62.3%	70.3%	833

4.2. Similar approaches

We present here the results given by similar approaches. We acknowledge that this is not a comparison because their work are based on different data and languages. (Lee et al., 2013) evaluate their system using the Ontonotes (Pradhan et al., 2011), a multilingual corpus, which contains 34.290 coreference chains, distributed in three languages: English, Chinese and Arabic.
(Garcia and Gamallo, 2014a) used a corpus (Garcia and Gamallo, 2014b) built from journalistic and encyclopedic texts, using the SemEval guidelines (Recasens et al., 2010). This corpus has annotations for Portuguese, Spanish and Galician. For Portuguese, this corpus includes texts from Portugal, Brazil, Mozambique and Angola, containing annotations of persons and pronouns.
In Table 2, we can see our results against the numbers reported by the authors of related work, all using the MUC (Vilain et al., 1995) metric. Note that this evaluation considers solely named entities (whereas our adapted system produces mixed chains, our gold reference contains only named entities).
Aware of the limitations of the analysis, we note that our results are well situated in the current state of the art, achieving a F-measure of 67.8% for all ten categories (70.3% when excluding time and value).
Garcia's system (Garcia and Gamallo, 2014a) which was evaluated on Portuguese but focused only on persons, achieved a F-measure of 87.4%. For Person named entities, our adapted model achieves 75.7% of F-measure, but it includes a lot more classes.

4.3. Manual evaluation and error analysis

For a more comprehensive evaluation, considering other NPs (besides named entities) we conducted a

Table 2: Evaluation of similar approaches - MUC metric (made on different corpora)

System	Lang-Categ	P	R	F
Lee	EN-NP-All	59.3%	62.8%	61.0%
Garcia	PT-NP-Pers	92.7%	82.7%	87.4%
Ours	PT-NE-All	81.0%	58.3%	67.8%
Ours	PT-NE-Pers	80.7%	71.3%	75.7%

manual analysis of the results. For that we considered a subset of the corpus, consisting of five texts, containing 25 coreference chains and 63 mentions. These chains include both named entities and common nouns. However, in this evaluation we could only calculate precision, since we don't have a reference. In this subset, our model presented 85.20% of precision.

Next, we present some errors which affected recall and precision of the model (when compared with the gold reference). As expected, the most common error occurs in the classes "Time" and "Value". The system loses links chains, such as: [18 km], [quase 18km] ([*almost 18km*]); [300.000km/s], [a velocidade da luz] ([*the light speed*]); [[o dia 05 de Janeiro] ([*the day 05 of January*]), [05/01]].

Other errors were due to our basic Acronym rule, as in the resulting chain - [P. tupynambai], [Paraguai], [Peru], [P.] - which groups togheter names refering to different entities, due to a possible common acronym. Since the system interpreted "P." as an acronym these noun phrases were grouped incorrectly.

There are also errors coming from the fact that the system does not rely on world knowledge, therefore it fails to recognize chains such as: [Ronaldo], [o Fenômeno] ([*the Phenomenon*]), which refer to name and nickname of a famous soccer player.

As a last example of error, the system sometimes links brands with their products, as in: [[Toyota Prius], [Prius], [A Toyota]]. In this example the car is in the same chain as the company, which is wrong.

5. Conclusion and Future Work

In this paper, we show our adaptation of Lee's approach for coreference resolution to Portuguese. We evaluate it considering ten named entity categories, and we show that these results are compatible with current state of the art. The system and related resources are available[1].

One main point to be considered in the future is to introduce semantic knowledge in our model, we plan to use Onto-PT(Oliveira and Gomes, 2014), a recent resource available for Portuguese. Through semantic relations is possible to identify implicit relations, as hyponymy, synonymy and hyperonymy, linking mentions like: [the bee], [the insect]; [the car], [the vehicle]; [the animal], [the dog].

[1] http://www.inf.pucrs.br/linatural/mpsfp.html

The inclusion of semantics is another reason for our option for a rule-based system. It seems that when knowledge in involved rule-based systems are a good option. (Hou et al., 2014) proposes a rule based system to solve bridging anaphora. As result, the authors show that their rule-based model outperforms learning-based approach, using the same knowledge resources. Also, learning approaches enriched with more features did not yield much improvement over the rule-based system.

Acknowledgments

The authors acknowledge the financial support of CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and FAPERGS (Fundação de Amparo à Pesquisa do Rio Grande do Sul).

6. References

Coreixas, T. (2010). Resolução de correferência e categorias de entidades nomeadas. Dissertação de Mestrado, Pontifícia Universidade Católica Do Rio Grande Do Sul.

Fernandes, E. R., dos Santos, C. N., and Milidiú, R. L. (2014). Latent trees for coreference resolution. Number 4, pages 801–835. Computational Linguistics - MIT Press.

Fonseca, E. B., Vieira, R., and Vanin, A. (2015). Dealing with imbalanced datasets for coreference resolution. In *Proceedings of The Twenty-Eighth International Flairs Conference - FLAIRS 2015*.

Freitas, C., Mota, C., Santos, D., Oliveira, H. G., and Carvalho, P. (2010). Second harem: Advancing the state of the art of named entity recognition in portuguese. In *Proceedings of Language Resources and Evaluation Conference*, Malta. - LREC 2010.

Garcia, M. and Gamallo, P. (2014a). An entity-centric coreference resolution system for person entities with rich linguistic information. pages 741–752.

Garcia, M. and Gamallo, P. (2014b). Multilingual corpora with coreferential annotation of person entities. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference - LREC 2014*, pages 3229–3233.

Hou, Y., Markert, K., and Strube, M. (2014). A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. pages 2082–2093. Proceedings of Conference on Empirical Methods on Natural Language Processing - EMNLP 2014.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. volume 39, pages 885–916. Computational Linguistics - MIT Press.

Martschat, S. and Strube, M. (2015). Latent structures for coreference resolution. *Transactions of the*

Association for Computational Linguistics, 3:405–418.

Oliveira, H. G. and Gomes, P. (2014). Eco and onto. pt: a flexible approach for creating a portuguese wordnet automatically. *Language Resources and Evaluation - LREC2014*, 48(2):373–393.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27. Association for Computational Linguistics.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8. Association for Computational Linguistics.

Silva, J. F. d. (2011). Resolução de correferência em múltiplos documentos utilizando aprendizado não supervisionado. Dissertação de Mestrado, Universidade de São Paulo.

Silva, W. D. C. (2013). Aprimorando o corretor gramatical cogroo. Dissertação de Mestrado, Universidade de São Paulo.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.