Improving Coreference Resolution with Semantic Knowledge

Evandro Fonseca[1(✉)], Renata Vieira[1], and Aline Vanin[2]

[1] Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil
evandro.fonseca@acad.pucrs.br, renata.vieira@pucrs.br
[2] Universidade Federal de Ciências da Saúde de Porto Alegre, Porto Alegre, Rio Grande do Sul, Brazil
aline.vanin@ymail.com

Abstract. This paper evaluates the impact of semantic features in coreference resolution for the Portuguese language. We show that the new proposed features obtained on the basis of currently available Portuguese semantic resources improve results in precision, recall and f-measure.

Keywords: Coreference resolution · Semantic knowledge · Information extraction · Machine learning

1 Introduction

The problem of coreference resolution has received a great deal of attention from the computational linguistics community. This problem usually requires previous language processing on many levels (POS-tagging, parsing, semantic analysis). Less resourceful languages may experience greater difficulty in advancing towards this kind of task. Usually, there is also a shortage of data. While, for English, we have the Ontonotes corpus [21], which contains 34.290 coreference chains, for Portuguese language we are only aware of the Summ-it corpus [5], which contains 560 coreference chains; the Garcia and Gamallo corpus [15] which is annotated only with regards to entities of type person; and the HAREM corpus [12] which contains named entities and their relations (including identity, which may considered as a form of coreference annotation). In spite of the smaller size, Summ-it is a high quality resource which has been used in many research initiatives ([6–8,11]). Regarding semantic resources, only recently we see large semantic databases as an alternative if one wants to go beyond the limits of the usual string matching and morphological heuristics for coreference resolution.

In this paper, we propose a model based on supervised machine learning for Portuguese coreference resolution, where semantic knowledge is considered in the elaboration of features. Our initial set of features is based on traditional previous work in the field, such as Soon et al.'s [25], enriched with features presented in more recent work such as Lee et al.'s [17], also following previous research on coreference for Portuguese [11]. On top of that we evaluate the impact of new features that makes use of semantic knowledge.

J. Silva et al. (Eds.): PROPOR 2016, LNAI 9727, pp. 213–224, 2016.
DOI: 10.1007/978-3-319-41552-9_21

The paper is organized as follows: Sect. 2 discusses the problem of coreference and the relevance of semantic knowledge for this problem; Sect. 3 presents related work; in Sect. 4, we describe the main Portuguese resources that we used in our system; Sect. 5 describes our model; in Sect. 6, we describe the experiments that measure the impact of the inclusion of semantic features; Sect. 7 presents our conclusions and future work.

2 Semantics for Coreference Resolution

Coreference resolution is an important task and also a great challenge for Natural Language Processing. It basically consists of finding different references to a same entity in a text, as in the example: [Schumacher] sofreu um acidente. [O ex-piloto] permanece em coma. ([*Schumacher*] *suffered an accident.* [*The ex-pilot*] *is still in coma*). In this case, the noun phrase [O ex-piloto] ([*The ex-pilot*]) is coreferent to [Schumacher].

There are cases where the coreference relation is simple to grasp, such as in [Barack Obama] and [Obama], in which both NPs share some identical part, in this case, "Obama". In other situations, establishing a coreference relation between two noun phrases is more complex. In cases such as [A abelha] and [O inseto] ([*The bee*] and [*The insect*]) there is a hyponymic relation between the two referents which is usually part of the readers' common sense. Besides, in this example, for Portuguese the two NPs differ in gender, a commonly used feature to deal with coreference. For a system to recognize this kind of relation, it would require an adequate semantic resource.

Although the Portuguese coreference resolution area is at an early stage of development, research in this task should be pursued, as it is quite relevant for many other tasks. In [13] it is shown that coreference resolution may provide meaningful gains for the area of entity relation extraction, since coreference links may be useful for extracting implicit relations. Consider the following sentence: [O presidente dos Estados Unidos], [Barack Obama], afirmou hoje que as alterações climáticas são a maior ameaça ao planeta. ([*The United States president*], [*Barack Obama*], *said today that the climate changes are a great threat for the planet*). When identifying and creating a coreference relation between [Barack Obama] and [o presidente] ([*the president*]), it is possible to infer a relation between the entities [Barack Obama] and [Estados Unidos] ([*United States*]) (in which Barack Obama is the president of the United States). In other words, when we say that Barack Obama is the president, it is possible to classify him as a person, as well as to say that he has a relation with the United States.

In general, previous coreference resolution systems are usually limited to non semantic lexical features, such as: string matching, heuristics for alias recognition, gender, number among others. We propose and evaluate a coreference resolution system for Portuguese with the addition of features based on entity categories and also world knowledge, considering semantic resources which were recently made available. Basically, our model combines known features, adapted from previous work with new semantic features based on semantic resources for

Portuguese. The development of more robust coreference resolution systems may help other NLP tasks for Portuguese, such as relation extraction and sentiment analysis.

3 Related Work

Coreference Resolution represents a great challenge in NLP, given the many levels of language that must be taken into account in this task. Analyzing the current state of the art, we see that it is hard to achieve good results, independent of the language. Most works go around without semantic resources, perhaps because the problem of domain independent semantic resources, that covers a broad range of semantic phenomena and that would fit into a particular processing need, was not solved yet.

One recent and relevant coreference resolution system is Lee et al.'s [17]. Their approach to coreference resolution combines the global information and precise features appointed 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 builds on the previous model's clusters output. In order to increase the recall, they combine several variations of matching.

Based on Lee's work, Garcia et al. [14] proposes Link-People: a model for coreference resolution which is tailored to person entities (what we may consider an initial semantic orientation). They considered three languages: Portuguese, Spanish and Galician. Their model combines the multi-pass architecture and a set of constraints and rules. The authors use some matching rules from [17]; in addition, they use a set of specific rules to dealing with pronouns, anaphora, cataphora for person entities. To detect person entities, the authors uses FreeLing NER [4]. In an error analysis, the authors mention the problem of lack of rich semantic resources, showing that their model could be improved by detecting semantic relations like synonymy, hyponymy and hyperonymy: [the boy] and [the youngster]. For Portuguese coreference, Garcia et al. built their own corpus [15] considering only entities of type person.

Although the semantic problem is mostly unattended, there is previous coreference resolution research that considers semantic knowledge for English. The authors of [22] evaluated the utility of world knowledge using a mention-pair and cluster-ranking model. For world knowledge, the authors used two knowledge bases: Yago and FrameNet. Their strategy consists into identifying relations like "Means", "IS-A" and "Type". Each relation is represented in YAGO as a triple. (AlbertEinstein, Type, physicist), for instance, denotes the fact that Albert Einstein is of type physicist. The relation "Means" provides different ways of expressing an entity, and therefore allows dealing with synonymy and ambiguity, i.e. for the two triples: (Einstein, Means, AlbertEinstein), and (Einstein, Means, AlfredEinstein) denotes the fact that Einstein may refer to the physicist Albert Einstein or the musician Alfred Einstein. From FrameNet, the authors used the semantic role related to verbs. For example: "Peter Anthony decries

the program trading as limiting the game to a few, but he is not sure whether he wants to denounce it because [...]". Note that the semantic role may help to establish a coreference link between "program trading" and "it", since with FrameNet it is possible to retrieve a relation between "decry" and "denounce", because these words appear in the same frame and the two noun phrases have the same semantic role. The authors show that each semantic source may offer some small gains but that their cumulative benefits can be substantial. This happens because coreference may be detected either through a semantic relation, like in [the boy] and [the youngster] connection, or through factual world knowledge, such as in [the United States president] and [Obama].

Dealing with English and deterministic rules (like Lee et al.'s), Hou et al. [16] proposes a rule based system to solve anaphora and bridging. Different from our work, which tries to identify coreference (identity relation), bridging resolution consists into recognizing and linking entities through non-identity relations. An example is the meronymyc relation ("Part_of") as in [the house] [the chimney]. To identify this type of relations, the authors used WordNet [19].

For Brazilian Portuguese, Silva [24] proposed a coreference resolution system based in the same Harem [12] semantic categories, using an unsupervised learning algorithm. To detect these categories, Silva used the parser PALAVRAS [1] and the named entities recognizer Rembrandt [3]. Regarding semantic processing, the author uses the synonymy relation based on Tep2.0 [18], a thesaurus containing synonymy and antonymy for Portuguese language. Silva reports that the semantic knowledge did not show improvements in his experiments. However, he considered a small corpus, containing just nine texts, which may be considered a limitation.

Coreixas [6] proposes a coreference resolution for Portuguese also focusing on the main categories of named entities: Person, Organization, Location, Work, Thing and Other. Resources and tools were the HAREM corpus [12], the parser PALAVRAS [1] and the Summ-it corpus [5]. In order to prove which of the semantic categories may help to solve coreferences, Coreixas compares two versions of her system. The author showed that the use of categories of entities resulted in an improvement in determining whether a pair is coreferent or not. Also, the importance of world knowledge for this line of research was mentioned, emphasizing the importance of databases with synonyms, such as Wordnet, to complement and support coreference resolution.

Following Coreixas, Fonseca et al. [11] proposes a machine learning system to solve coreference in Portuguese but considering only proper names. To detect named entity categories, the authors used resources such as Repentino [23] and NERP-CRF [9], plus auxiliary lists containing common nouns, referring to certain categories, such as professions "advogado, agrônomo, juiz..." (*lawyer, agronomist, judge...*) for person and "avenida, rua, praça, cidade..." (*avenue, street, square, city...*) for location.

We see that the previous work combines the use of named entity categories and semantic knowledge resources. However, the only Portuguese semantic resource considered for this task was Tep2.0, which contains 8.528 synonym and

antonym relations. There are currently available more comprehensive semantic data bases. Onto-PT [20], contains 168.858 synonymy relations, 91.466 hyperonymy/hyponymy, 9.436 meronymy and 92.598 antonymy relations.

4 Corpus and Semantic Resources

In this section, we cite the main resources used to build our coreference resolution system: the coreference annotated corpus and semantic resources.

4.1 Summ-it Corpus

Summ-it [5] is a corpus consisting of fifty journalistic texts from the Science section of Folha de São Paulo newspaper. It is part of the PLN-BR corpus [2]. The texts were annotated with syntactic, coreference and rhetorical structure information. Summ-it also includes summaries constructed manually and automatically. The corpus has a total of 560 coreference chains with an average of 3 mentions (noun phrases for each chain). The largest chain has 16 mentions. Summ-it has been used in previous coreference resolution research for Portuguese ([6–8,11]) and has had an important role in the training and validation of classification models. The coreference annotation scheme of Summ-it is distributed in tree distinct files: Markables, Words and POS.

A Markables file indicates the mentions. Each markable contains an id, a pointer to the corresponding set of words (span), a link to a coreference set (member), status (new or old) and type of NP (definite, indefinite, pronoun, etc.), for example:

– < *markable id*= "markable_95" span="word_23..word_25" member="set_2" status="new" np_n="yes" np_form="def-np" / >

In this example, the element "span" indicates word ids (23 to 25), that correspond to a reference to the Words file, which contains the list of tokens. As in the example below, they refer to the noun phrase [o agrônomo Miguel Guerra] ([*the agronomist Miguel Guerra*]).

– < *word id* ="*word_23*"> *o* < /*word* >
< *word id* ="*word_24*"> *agrônomo* < /*word* >
< *word id* ="*word_25*"> *Miguel_Guerra* < /*word* >

Summ-it also contains POS files, which contains the word id, the grammatical class of the word, gender and number (singular/plural), as seen below:

– < *word id* ="*word_25*"><*prop canon* ="*Miguel_Guerra*" *gender* = "*M*" *number* ="*S*"><*secondary_prop tag* ="*hum*"/ ><*/prop* >
< /*word* >

4.2 Semantic Resources

As semantic resources we used the semantic annotation provided by the Palavras parser [1] and Onto-PT [20]. From Palavras we used the semantic tags for identifying references to person, organization and location. Onto-PT was considered as a general semantic base. Similarly to WordNet [19], it is structured in synsets (groups of synonymous word senses that can be seen as possible lexicalizations of a natural language concept) and semantic relations connecting synsets, including not only hyperonymy (a concept is a kind of another) and part-of (a concept is part of another), but also others, such as causation (a concept causes another) or purpose-of (a concept is used for another). To extract the relations, we utilized Onto-PT API which, for a given pair of words, retrieves all relations between the given elements, as in the examples given in Table 1. Although Onto-PT has several relations, in this work, we focused only in hyponymy, hyperonymy and synonymy relations.

Table 1. Onto-PT: examples of semantic relations returned for a pair of words.

Word pairs	Relations
estudo, pesquisa *study, research*	sinonimoDe/synonymOf
abelha, inseto *bee, insect*	hiponimoDe/hyponymOf
animal, cachorro *animal, dog*	hiperonimoDe/hyperonymOf

5 Enriched Semantic Model

Our model is based on features inspired by Lee et al.'s [17] and Soon et al. [25]. We converted some Lee et al.'s rules into features, such as: Relaxed_String_Match and Word_Inclusion; and adopted some Soon et al.'s features, like Number, Gender and Alias (which have been widely used by many related work). Some features were simply reimplemented, others had to be adapted. In special, to these initial features, we added five new semantic features: two based on entity categories (person, place and organization) and three using Onto-PT. To provide the candidate-pairs, we use the same strategy of Fonseca et al.'s [10], each noun phrase makes pair with their followers. In order to balance the data set, we choose randomly n negative pairs (where n is the quantity of positive pairs). Then we run one hundred times a ten fold cross validation for each model.

5.1 Basic Features

Next we present the basic (non-semantic) features of our model. The first six features were straightforward re-implementations, while features 7 to 12 required some form of adaptation for Portuguese, as described below.

1. Exact String Match: if the current NP and antecedent are equal.
2. Relaxed String Match: if the strings up to the head nouns are equal.
3. Word Inclusion: if there are no different words (nouns, verbs, adjectives, adverbs) in the NP, when compared to the antecedent.
4. Alias: if a NP is acronym of the other.
5. NP Distance: The distance of two NPs is given in terms of the number of NPs between them.
6. Sentence Distance: This feature explores a similar idea but now we count the distance in number of sentences.
7. Embedded Nps: This feature explores the NPs structure, checking if a noun phrase is not identical to a constituent part of the other. To recognize constituents, we observe the presence of prepositions like "de" and "em". In (a), as the noun phrases are embedded, they could no be coreferent.

   (a) [O garoto da casa ao lado] ... [a casa ao lado] ([*The boy of the next house*]) ... ([*the next house*]).
8. Proper Noun Word Match: This feature returns true if three conditions are satisfied[1]: both noun phrases must contain proper nouns; the proper nouns must be equal; and these NPs are not embedded. Example (b) shows a violation of the third condition.

   (b) [Califórnia]... [região sul da Califórnia] ([*the southern of California*]).
9. Gender: Nouns in Portuguese have gender. While in English the NP [the teacher] may link to female or male name, in Portuguese, we would have either [o professor] or [a professora]. But this may be tricky: [a abelha] and [o inseto] ([*the bee*] and [*the insect*]) with different genders, may be coreferent.
10. Number: If the phrases agree in number. Here we have a difference in the articles. For Portuguese, for example, we have [os professores], where both noun and article have plural forms. Like in the Gender feature, we may have coreference links between NPs different in number, as in (c):
    (c) [Um fóssil de tiranossauro Rex] foi encontrado no oceano Atlântico. [Os ossos]... ([*A fossil of tyrannosaur Rex*] *was found in Atlantic Ocean.* [*The bones*])...

5.2 Semantic Features

To these previous features, we added five semantic features: Entity Category Equal, Entity Category Different, Hyponymy, Hyperonymy and Synonymy:

11. Entity Category Equal: This feature sees if the NPs have the same semantic category. To extract these categories we use the parser PALAVRAS [1]. We consider Person, Location and Organization semantic categories.
    (d) A opinião é de [Miguel Guerra], da UFSC (Universidade Federal de Santa Catarina). Para [o agrônomo] ... (*The opinion is from* [*Miguel Guerra*],

[1] Different from [17], we did not implement rules "location and numeric mismatches".

of UFSC (Universidade Federal de Santa Catarina). To [*the agronomist*] ... In text fragment (d), the pairs of NPs [Miguel Guerra] and [o agrônomo] ([*the agronomist*]) are both NPs of the category "Person". As not all NPs have a category associated to it, we consider another feature, "Entity Category Different";

12. Entity Category Different: This feature sees whether the semantic categories of the NPs are different. This feature is used here for the cases where there is no category associated to an NP, in which case both features: "Entity Category Equal" and "Entity Category Different" return false.
13. Hyponymy: This feature basically extracts the lemma from the noun phrases head word and check if they are in hyponymy relation in OntoPT. This feature helps to identify relations as in (e):
    (e) Já se perguntou como [as abelhas] fabricam mel? [Os insetos] saem em busca de... (*Ever wonder how* [*the bees*] *make honey?* [*The insects*] *seek out* ...)
    To avoid the incorrect links (f), we combine the pre and post modifier techniques in this feature.
    (f) Foi o tempo em que decifrar [o genoma].... [o quebra-cabeça genético]... Isso é [um problema ambiental]... (*There was a time in which to decipher* [*the genome*] ... [*the genetic puzzle*] ... *This is* [*an environment problem*] ...)
    In "f" [o quebra-cabeça genético] ([*the genetic puzzle*]) is not coreferent with [um problema ambiental] ([*an environment problem*]). But if we analyze just the relation between their head words "puzzle" and "problem", the semantic relation is found. In other words, the feature Hyponymy returns "true", if the following constraints are satisfied:
    – the lemma of head words must be in an hyponymy relation;
    – the Word_Inclusion feature must return "false".
14. Hyperonymy: Similar to Hyponymy, this feature extracts the lemma from the noun phrases head word and search for their hypernymy relation in OntoPT. Like Hyponymy, this feature also combines pre and post modifier restrictions.
15. Synonymy: Like before, the synonymy relation is verified in OntoPT and pre and post modifier restrictions are considered. This feature helps to link mentions as in (g).
    (g) O trabalho de pesquisadores da USP está revelando uma série de novas espécies de [um tipo especial de fungo]. [Pequenos cogumelos]... (*The work of researchers from USP is revealing a number of new species of* [*a special type of fungus*]. [*Small mushrooms*]...)

6 Experiments

Next we describe our experiments in which six models are evaluated. The Baseline considers only the non semantic features. The second model, EntityCat, adds the features related to entity categories to the Baseline. The next model adds

the Synonymy feature to EntityCat. Then we add Hyponymy to EntityCat. The fifth model adds Hyperonymy to EntityCat and, finally, the last model includes all the five semantic features.

We extracted the candidate pairs, initially retrieving 3022 positive pairs and 94889 negatives. In order to generate a balanced model, we utilized the random undersampling technique, which consists into choosing randomly a sample from negative pairs, the same number of positive samples (3022). This method was used in [10], presenting satisfactory results. To build the classifiers, we used J48 implementation with ten-fold cross validation. To compare the models we built one hundred versions of each (due to the randomly undersample) and present the resulting average in Table 2.

The data was submitted to a Tukey's test, using 5 % of probability. We can see that the all variations including semantic features presented a significant improvement in recall for the positive class, without loss of precision. Note that the Full Semantic model or the EntityCat+Hyponymy are always better than some of the other models in all measures and classes. Perhaps synonymy and hyperonymy are not so influential, due to the fact that the quantity of positive examples involving these features was small in our data set, so that the learning algorithm do not consider it relevant.

Table 2. Semantic models evaluation

Model	Avg Prec Pos	Avg Prec Neg	Avg Recall Pos	Avg Recall Neg	Avg F-Pos	Avg F-Neg
Baseline	79.16 % abc	64.99 % b	53.81 % b	85.66 % a	63.96 % b	73.87 % c
EntityCat	78.57 % c	66.37 % a	57.26 % a	84.24 % b	66.16 % a	74.21 % b
EntityCat +Synonymy	78.61 % c	66.36 % a	57.20 % a	84.30 % b	66.13 % a	74.22 % b
EntityCat +Hyponymy	79.73 % ab	66.59 % a	57.13 % a	85.41 % a	66.52 % a	74.82 % a
EntityCat +Hyperonymy	79.04 % bc	66.36 % a	56.99 % a	84.73 % ab	66.13 % a	74.39 % b
Full Semantic	79.92 % a	66.56 % a	56.95 % a	85.62 % a	66.45 % a	74.88 % a
CV	2.68 %	1.17 %	4.65 %	3.12 %	1.85 %	1.06 %
σ	2.13	0.77	2.63	2.65	1.22	0.79

When the numbers across the same collumn are followed by the same letters, their difference is not significant (accorging to the Tukey's test ($p < 0.05$)). Contrasted along the vertical line, 'a' means a better result than 'b', and 'b' than 'c' (a>b>c). '*CV*' represents the coefficient of variation and 'σ', the standard deviation.

In general, Table 2 shows that the semantic knowledge improves the coreference resolution in several aspects, allowing the generation of coreference chains which are not only based on morpho-syntatctic and string matching features. Another point to consider is that the occurrence of coreference links between semantic related NPs (without string similarity) is less frequent, so it is not a surprise that the improvements are not expressed in larger numbers.

In other words, with the semantic model, it was possible to identify a few new, but more interesting and difficult cases of coreference relations, such as:

[o fungo], [pequenos cogumelos] ([*the fungus*], [*small mushrooms*]); [o álcool], [o etanol] ([*the alcohol*], [*the ethanol*]).

7 Conclusion

In this paper, we evaluated the impact of adding semantic features to a Portuguese coreference resolution system. The semantic features are based on the identification of entity categories (person, place and organization) and on semantic relations of synonymy, hypernymy and hyponymy which are provided by the Portuguese semantic resource, Onto-PT. As a result, we show improvements in the Portuguese coreference resolution task. These semantic features allow the identification of coreferent pairs which share semantic similarity that are not realized through lexical similarity. Although we found apparently just a little improvement, it certainly adds quality on a new level to the output. As future work, we want to test our semantic model at the chains level, using the CoNLL scorer [21]. The source files used to generate the features are available to the community.[2]

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).

References

1. Bick, E.: The parsing system "palavras": automatic grammatical analysis of Portuguese in a constraint grammar framework. Ph.D. thesis, Aarhus University, Aarhus University Press, Denmark (2000)
2. Bruckschen, M., Muniz, F., Souza, J., Fuchs, J., Infante, K., Muniz, M., Gonçalves, P., Vieira, R., Aluísio, S.: Anotação linguística em xml do corpus pln-br. Série de relatórios do NILC (2008)
3. Cardoso, N.: Rembrandt - a named-entity recognition framework. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation - LREC, Istanbul, Turkey, pp. 1240–1243 (2012)
4. Carreras, X., Màrquez, L., Padró, L.: A simple named entity extractor using adaboost. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, vol. 4, pp. 152–155. Association for Computational Linguistics (2003)
5. Collovini, S., Carbonel, T.I., Fuchs, J.T., Coelho, J.C., Rino, L., Vieira, R.: Summ-it: Um corpus anotado com informações discursivas visando a sumarização automática. In: Proceedings of V Workshop em Tecnologia da Informação e da Linguagem Humana, Rio de Janeiro, RJ, Brasil, pp. 1605–1614 (2007)
6. Coreixas, T.: Resolução de correferência e categorias de entidades nomeadas. Pontifícia Universidade Católica Do Rio Grande Do Sul, Dissertação de Mestrado (2010)

[2] http://www.inf.pucrs.br/linatural/scorref.html.

7. da Silva, F.J.V., Carvalho, A.M.B.R., Roman, N.T.: A comparative analysis of centering-based algorithms for pronoun resolution in Portuguese. In: Kuri-Morales, A., Simari, G.R. (eds.) IBERAMIA 2010. LNCS, vol. 6433, pp. 336–345. Springer, Heidelberg (2010)
8. de Souza, J.G.C., Gonçalves, P.N., Vieira, R.: Learning coreference resolution for Portuguese texts. In: Teixeira, A., de Lima, V.L.S., de Oliveira, L.C., Quaresma, P. (eds.) PROPOR 2008. LNCS (LNAI), vol. 5190, pp. 153–162. Springer, Heidelberg (2008)
9. do Amaral, D.O.F.: O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa. Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul (2013)
10. Fonseca, E.B., Vieira, R., Vanin, A.: Dealing with imbalanced datasets for coreference resolution. In: Proceedings of The Twenty-Eighth International Flairs Conference - FLAIRS (2015)
11. Fonseca, E.B., Vieira, R., Vanin, A.A.: Coreference resolution in portuguese: detecting person, location and organization. J. Braz. Comput. Intell. Soc. **12**, 86–97 (2014)
12. Freitas, C., Mota, C., Santos, D., Oliveira, H.G., Carvalho, P.: Second HAREM: advancing the state of the art of named entity recognition in Portuguese. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC, Valletta, Malta (2010)
13. Gabbard, R., Freedman, M., Weischedel, R.: Coreference for learning to extract relations: yes, virginia, coreference matters. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, vol. 2, pp. 288–293. Association for Computational Linguistics (2011)
14. Garcia, M., Gamallo, P.: An entity-centric coreference resolution system for person entities with rich linguistic information. In: Proceedings of 25th International Conference on Computational Linguistics, COLING, Dublin, Ireland, pp. 741–752 (2014)
15. Garcia, M., Gamallo, P.: Multilingual corpora with coreferential annotation of person entities. In: Proceedings of the 9th edn. of the Language Resources and Evaluation Conference - LREC, pp. 3229–3233 (2014)
16. Hou, Y., Markert, K., Strube, M.: A rule-based system for unrestricted bridging resolution: recognizing bridging anaphora and finding links to antecedents. In: Proceedings of Conference on Empirical Methods in Natural Language Processing - EMNLP, Doha, Qatar, pp. 2082–2093 (2014)
17. Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic coreference resolution based on entity-centric, precision-ranked rules. Comput. Linguist. **39**, 885–916 (2013). MIT Press
18. Maziero, E.G., Pardo, T.A., Di Felippo, A., Dias da Silva, B.C.: A base de dados lexical e a interface web do tep 2.0: thesaurus eletrônico para o português do brasil. In: Proceedings of the XIV Brazilian Symposium on Multimedia and the Web, pp. 390–392. ACM (2008)
19. Miller, G.A.: WordNet: a lexical database for english. Commun. ACM **38**(11), 39–41 (1995)
20. Oliveira, H.G., Gomes, P.: ECO and Onto-PT: a flexible approach for creating a portuguese wordnet automatically. Lang. Resour. Eval. **48**(2), 373–393 (2014)

21. Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., Xue, N.: Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, pp. 1–27. Association for Computational Linguistics (2011)
22. Rahman, A., Ng, V.: Coreference resolution with world knowledge. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 814–824 (2011)
23. Sarmento, L., Pinto, A.S., Cabral, L.M.: REPENTINO – A wide-scope gazetteer for entity recognition in Portuguese. In: Vieira, R., Quaresma, P., Nunes, M.G.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.) PROPOR 2006. LNCS (LNAI), vol. 3960, pp. 31–40. Springer, Heidelberg (2006)
24. da Silva, J.F.: Resolução de correferência em múltiplos documentos utilizando aprendizado não supervisionado. Dissertação de Mestrado, Universidade de São Paulo (2011)
25. Soon, W.M., Ng, H.T., Lim, C.Y.: A machine learning approach to coreference resolution of noun phrases. Comput. Linguist. **27**(4), 521–544 (2001)