Dealing With Imbalanced Datasets For Coreference Resolution

[1]Evandro B. Fonseca, [1]Renata Vieira, [2]Aline A. Vanin

[1]Pontifícia Universidade Católica do Rio Grande do Sul
Avenida Ipiranga, 6681 – Faculdade de informática
CEP 90619-900 Porto Alegre RS, Brasil

[2]Universidade Federal de Ciências da Saúde de Porto Alegre
Rua Sarmento Leite, 245 – Departamento de Educação e Humanidades
CEP 90050-170 Porto Alegre RS, Brasil

e-mails: evandro.fonseca@acad.pucrs.br, renata.vieira@pucrs.br, aline.vanin@ymail.com

Abstract

In this paper we present our proposed model for coreference resolution and we discuss the imbalanced dataset problem related to this task.We conduct a few experiments showing how well our set of features can solve coreference for Portuguese. In order to minimize the imbalance between the classes, we evalaluated the system on the basis of well known re-sampling techniques.

1. Introduction

In this paper, we propose a supervised machine learning model for Portuguese coreference resolution. The features are based on Lee's rules (Lee et al., 2013).. Like many machine learning works, in our coreference resolution model, the data set presents an imbalanced distribution over the classes. This is an interdisciplinary and well-known problem: as we can see in Chawla et al. (2002), Akbani et al., (2004) and Cieslak et al. (2008). In our study, we aim at analyzing whether the undersampling techniques help to improve coreference resolution over the imbalanced dataset that we use to train our proposed model.

The rest of this paper is organized as follows: Section 2 introduces the notion of coreference; Section 3 presents the problem of imbalanced classes in coreference resolution; Section 4 presents related work about imbalanced datasets and coreference resolution; in Section 5, we describe our proposed model; in Section 6, we describe our experiments involving different balancing levels; in Section 7, the conclusions and the future works are presented.

2. Coreference Resolution

Coreference resolution is a process which, basically, consists into finding different references of a same entity in a text, as in the example: (1) "Schumacher$_{[i]}$ sofreu um acidente. O ex-piloto$_{[i]}$ permanece em coma". ["Schumacher$_{[i]}$ suffered an accident. The ex-pilot$_{[i]}$ is still in coma"]. In this case, the noun phrase "O ex-piloto" ["The ex-pilot"] is a coreference of "Schumacher". Coreference resolution is a relevant task and a great challenge for computational linguistics. While it is relatively easy to grasp coreference relations such as (2) "Jeff Mills" and (3) "Mills", in which both NPs carry part of the noun "Mills", it is a very complex task to relate the following noun phrases: (4) "A abelha" [The bee] and (5) "O inseto" [The insect]. As there is a hyponymic relation between "the bee" and "the insect", it would require the use of an accurate ontology to deal with such cases. Besides, in Portuguese, the gender is different in each NP: in (4), the head of the NP is feminine and, in (5), it is masculine. When dealing with this language, this challenge is even harder, because the quantity of resources is limited when compared to other languages, such as English. The lack of resources for Portuguese may be seen, for example, in this comparison: Ontonotes (Pradhan et al., 2011) is a corpus for English language with around 1.3 million of words, distributed in five layers of annotation: Syntactic layer, Propositional layer, Named Entities layer, Word Sense layer and Coreference layer. For coreference resolution, there is a total of 131,886 mentions, 97,556

links and 34290 chains; Harem corpus (Freitas, 2010), for Portuguese, has around 225 thousand words, distributed in three layers: Coreference layer, Relation between Named Entities layer (4803 marks) and Semantic Category layer (7847 recognized named entities).

Although it is at an early stage of development, the research for coreference resolution in Portuguese should be pursued, as it is quite relevant for many other tasks. Gabbard et al. (2011) show that coreference resolution may provide meaningful gains for the relation extraction among named entities, since the coreference links may be useful for extracting sets of implicit relations. Consider the following sentence: (6) "José da Silva mora perto do Centro, em Porto Alegre. O aluno está no primeiro ano de seu mestrado na PUCRS". [José da Silva lives close to the Downtown area. The student is taking his first year of Masters at PUCRS]. When identifying and creating a coreference relation between "José da Silva" and "student", it is possible to infer a direct relation between the entities "José da Silva" and "PUCRS" (in which José da Silva is a student at PUCRS). In other words, when we say that José da Silva is a student, it is possible to classify him as a person, as well as to say that he has relation with PUCRS.

Research on relation extraction is being conducted for Portuguese (Abreu et al. ,2013). In this work the recognition of relations among named entities such as Person, Location and Organization could benefit from our research on coreference resolution.

3. The Imbalanced Class Problem in Coreference Resolution

The imbalance class problem typically occurs when, in a classification problem, there are many more instances of some classes than others. In such cases, the standard classifiers tend to favor the majority class ignoring the small ones. In practical applications, the number of imbalanced instances can be drastic, such as 1 to 100, 1 to 1000 or 1 to 10000 (sometimes even more) (Chawla et al., 2004). This problem is prevalent in many applications, including: fraud/intrusion detection, risk management, text classification, and medical diagnosis/monitoring and others. For the coreference resolution, for example, this proportion also exists. Given the text fragment: "A opinião é do agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina), [...]" ["The agronomist Miguel Guerra's opinion, from UFSC (Federal University of Santa Catarina), […]"], we must cross all noun phrases, generating the candidate pairs, as in Table 1:

Table 1: Candidate pairs.

NP(x)	NP(y)	Coreference
o agrônomo	Miguel Guerra	Yes
o agrônomo	a UFSC	No
o agrônomo	Universidade Federal de Santa Catarina	No
Miguel Guerra	a UFSC	No
Miguel Guerra	Universidade Federal de Santa Catarina	No
UFSC	Universidade Federal de Santa Catarina	Yes

Each noun phrase forms a pair with the next one, but never with the previous one, like the method used by Fonseca et al., (2013). Note that the quantity of negative examples in this little text fragment is 2 negative pairs for each positive one. Using the Summ-it corpus (Collovini et al., 2007) to build our supervised coreference dataset, this proportion has achieved 31 to 1.

Learning in presence of imbalanced datasets is an important issue in machine learning. Learning algorithms incorporate the assumption that maximizing the overall accuracy is the goal. However, in many cases this does not meet the goals and requirements of an application and the results are unsatisfactory classifiers. In coreference resolution, for example, we prioritize the positive class. In other words, we try to minimize the false positives. A common way to deal with imbalanced data is to artificially change the data set distribution by oversampling or/and undersampling. Oversampling consists in replicating some of the training examples of the minority class until the desired class distribution is reached..While the undersampling consists in removing some training examples of the majority class. There are many techniques for selecting the training examples that should be replicated or deleted. The most popular method is random selection. Both techniques have correlated problems: undersampling can eliminate potentially useful examples of data, while the oversampling may increase the probability of occurrence of an overfitting, since the most accurate oversampling methods are examples of the minority class copies. In this sense, a symbolic model classification, for example, could build rules that are seemingly accurate, but actually cover replicated examples.

Some recent researches are focused on trying to overcome these problems for these two classes of methods. An example of this is Chawla's research (2002), which combines sampling techniques, and, instead of a simple oversampling replication of the minority class, they do the interpolation of pixels (in this case, the interpolation consists into a randomly replication of the "x" neighbors from each sample). Thus, the overfitting is minimized and the boundaries of decision to the minority class are extended.

Another technique, which may improve the balance among the classes, is the cost-sensitive learning: the training examples assign relative costs, called misclassification costs. The idea is to set high costs for the examples of the minority class and lower costs for the examples of the majority class. The goal of learning algorithms is to minimize the total costs that would result from classifying the training examples. This approach requires a cost-sensitive implementation of a particular learning algorithm. However, appropriate oversampling can have the same effect (Cimiano, 2006).

The variation of threshold may be an efficient technique, too: the method consists into varying the classifier threshold. Internally, learning algorithms estimate the probability that an example is assigned to each class (probabilistic class distribution). Basically, the class having the highest probability is normally assigned. Thus, a binary classifier chooses that class that has a probability greater than 0.5. If this value is increased for the majority class, the threshold to assign the majority class is higher and the classifier will probably assign the minority class. Sebastiani et al. (2002) perform a distinction between the forms to derive a value for the threshold analytically (through known metrics) and experimentally (tuning the threshold).

In our study we evaluate the impact of the undersampling technique, avoiding the effect of overfitting caused by the oversampling.

4. Related work

In this section we describe related works that deal with imbalanced dataset and coreference resolution.

4.1 Imbalanced Dataset

Japkowicz (2000) discusses the effect of imbalance in a dataset. The author has evaluated three strategies: Resampling, Down-Sizing and Learning by recognition. In Resampling, two resampling methods were considered in this category. Random resampling consists in resampling the smaller class at random until it consists of as many samples as the majority class, and focused resampling consists of resampling only those minority examples that occur on the boundary between the minority and majority classes. Down-Sizing consists in two down-sizing methods, closely related to the re-sampling methods. The first one, rand_downsize, consists of eliminating randomly elements of the over-sized class until it matches the size of the minority class. The second one, focused_downsize, consists of eliminating only further away elements. Learning by Recognition strategy consists of using a trained Multilayer Perceptron algorithm, in order to auto associate the samples. Japkowicz noted that both the sampling approaches were effective, and she also observed that using the sophisticated sampling techniques did not give any clear advantage in the considered domain.

Estabrooks et al. (2004) propose experiments involving different levels of undersampling and oversampling. In this research, they concluded that neither the oversampling nor undersampling strategies are always the best ones to use. Finding a way to combine them could – perhaps – be useful, especially if the bias resulting from each strategy has a different nature.

Chawla et al. (2002) proposes a method called SMOTE (Synthetic Minority Oversampling Technique). This technique consists of creating new instances through interpolation. For each positive instance, its nearest positive neighbors were identified and new positive instances were created and placed randomly in between the instance and its neighbors.

Ling and Li (1998) have combined over-sampling of the minority class with under-sampling of the majority class, showing that the over and undersampling techniques combination did not provide a significant improvement.

In this paper, we describe experiments involving resampling techniques, aiming to increase the performance of our coreference resolution model.

4.2 Coreference Resolution:

Coreference resolution is a well-known NLP problem. In the literature, we find works that are based only in rules, and others that are based only on machine learning. At the Conference on Computational Natural Language Learning (CoNLL 2011), Lee et al. (2011) presented a system that is purely based on rules for coreference resolution in English. The *Sanforfd's Multi-Pass Sieve Coreference Resolution System*, which is purely deterministic, reached an efficiency of 57.79%, and this was measured by the average rate among three performance metrics (MUC, B-CUBED and CEAFe), described in Pradhan et al. (2011). In his most recent publication, Lee et al. (2013) present a system in a more specific way, describing the 10 Sieves. Each sieve links the mentions until that specific rules be satisfied.

Coreixas (2010) proposes a coreference resolution system for Portuguese (PT-BR), with focuses on categories of named entities. According to the author, works related to English have had successful results when using specific categories of entities. Based on these assumptions, Coreixas hypothesized that the use of specific categories of named entities has a positive impact on the task of coreference resolution, since each category has distinct and well-defined characteristics. As the categorization defines the field, the use of semantic information as a support tool in the coreference resolution process becomes more feasible. Coreixas' system was based on machine learning, categorization of named entities, such as Person, Organization, Location, Work, Thing and Other (from the corpus of

HAREM (Freitas et al., 2010)), the parser PALAVRAS (Bick, 2000) and the resource from Summ-it corpus (Collovini et al., 2007). Coreixas compares two versions of the system, namely: Baseline and "Recorcaten" (REsolução de CORreferência por CATegorias de ENs, meaning coreference resolution by named entities categories). The first version aimed at generating pairs of phrases without considering the categories of NEs. The second generates pairs considering these types of entities. As a contribution, through experiments with both versions, Coreixas showed that the use of categories of entities provided an improvement in the percentage of correct answers to determine whether a pair is anaphoric or not. Also, it showed the importance of world knowledge for this line of research, given the fact that some categories, such as Event and Organization, did not show a satisfactory return on the classification of coreferent pairs. This happens because the process of disambiguation was not performed correctly, thus emphasizing the importance of databases with synonyms, such as Wordnet (Miller, 1995), to complement and support the resolution of coreference. This work has limitations: (a) the size of the corpus used in the experiments is not very big (it has only fifty texts); (b) there are not many resources for Portuguese; (c) according to Coreixas, the parser presents several problems of annotation.

Fonseca et al. (2013) proposes a coreference resolution system for Portuguese (PT-BR) using supervised machine learning. Their system has a total of 9 features. The author has focused only in pairs whose NPs were proper nouns,.
In our current work we combine features previously proposed in Fonseca et al. (2013) and Lee et al.'s (2013) established rules. In this work we consider all noun phrases pairs..To study the imbalance problem, we ran a few experiments, aiming to find the best balance for our algorithm.

5. The Proposed Model

This section describes the resources and the set of implemented features. The resources used in the construction of our coreference resolution model are: Corpus Summ-it (Collovini et al., 2007); the Portugues parser PALAVRAS,(Bick, 2000) for named entity recognition; and Weka (Boukckaert et al., 2013), a collection of machine learning algorithms.

We implemented twelve features, as follows below. They are an adaptation for Portuguese of Lee et al. (2013)'s along with features from Fonseca et al. (2013).
(1)Exact_String_Match: If the two NPs are exactly the same, Ex: [The agronomist] [The agronomist]; **(2) Relaxed_String_Match:** If the strings obtained by dropping the text following the heads are identical. Ex: [The museum of Porto Alegre]... [the museum]; **(3) Word_Inclusion:** If NP2 has the same modifiers of NP1 or has a subset of the NP1 modifiers. Ex: [the correct runway]… [the wrong runway], in this case, returning false; **(4) Not_IwithinI:** If a mention is not in an IwithinI construction. Ex: for [The boy with a brown t-shirt] …[the brown t-shirt], the feature returns false, because the NPs are in IwithinI construction; **(5) Proper_Head_Word_Match:** This feature returns true if three conditions are satisfied: Both NPs must be proper nouns, the heads of the NPs must share some same elements and these NPs are not in IwithinI construction Ex: [Adalberto Portugal] and [Portugual] for this case, this feature return false, because the NPs is in a IwithinI construction ; **(6) Alias:** If one of the words from NP1 is acronym of NP2; **(7) Gender:** If the phrases agree in gender (male/female); **(8) Number:** If the phrases agree in number (singular/plural); **(9) Semantic_Categ_Eq:** If the categories of entities (Person, Location or Organization) are equal; **(10) Semantic_Categ_Dif:** If the categories (Person, Location or Organization) are different. When the category is unknown, both Semantic_Categ_Eq and Semantic_Categ_Dif are *false*; **(11) *NP_Distance:*** The distance between NP1 NP2 in number of NPs **(12) Sentence_Distance:** The number of sentences between NP1 and NP2 is counted.

The construction of NP pairs was based on coreference information contained in the Summ-it (Collovini et al., 2007) corpus. To create the pairwise, we use Fonseca et al.'s (2013) strategy: each noun phrase makes pair with the next one, never with the previous one. In order to provide the semantic categories, we use PALAVRAS (Bick, 2000). Using the Summ-it corpus as input, we obtained a dataset consisting of 3022 coreferent pairs and 94889 non-coreferent pairs. If we divide these results in the original dataset, we have 31 negative pairs for each positive one (94889/3022= ~31). We decided to run this imbalanced dataset on Weka API (Boukckaert et al., 2013) in order to create a baseline margin. For this and other experiments, we use the J48 decision tree. We chose this algorithm because in balancing tests it presented the best results. The results can be seen in the confusion matrix (Table 2):

Table 2: Confusion Matrix of J48.

-	**P**	**N**
P	927	2095
N	321	94568

Table 3: Precision, recall and f-measure for Baseline model, using J48 algorithm.			
-	**Precision**	**Recall**	**F-measure**
P	74.3%	30.7%	43.4%
N	97.8%	99.7%	98.7%

In Table 3, we can see that the negative class was privileged because the supervised machine learning algorithms aim at improving the global accuracy, which is 97.5%. However, this is not a good result. In coreference resolution, using machine learning, we aim at improving the positive class, in addition to reducing the false positives. In this context, the negative class is not so important. Analyzing Table 2, note that, from 3022 positive pairs, we had 2095 false negatives (decreasing the recall of positive class) and 321 false positives (decreasing its precision).

6. Experiments

Aiming to improve the results from positive class, a few experiments were conducted. Each balancing experiment was run one hundred times, in order to calculate an average from precision, recall and f-measure. We did five experiments, choosing randomly a different number of samples from class "N", as in Tables 4 and 5.

Table 4: Experiment results from P class.			
Number of P and N samples	Avg. Precision	Avg. Recall	Avg. F-measure
3022 – 3022	79.5%	56.0%	65.6%
3022 – 6044	90.6%	40.1%	55.6%
3022 – 9066	92.0%	37.7%	53.4%
3022 – 15110	89.9%	36.6%	52.0%
3022 – 30220	85.6%	35.3%	50.0%

Table 5: Experiment results from N class.			
Number of P and N samples	Avg. Precision	Avg. Recall	Avg. F-measure
3022 – 3022	66.0%	85.4%	74.4%
3022 – 6044	76.6%	98.0%	85.9%
3022 – 9066	82.6%	98.9%	90.0%
3022 – 15110	88.7%	99.2%	93.6%
3022 – 30220	93.9%	99.4%	96.6%

In Tables 4 and 5, respectively, we can see the results from positive and negative classes. The quantity of positive pairs extracted from Summ-it is 3022 (we use all positive pairs). The class "N" has a total of 94889 pairs, suffering undersampling in all experiments. Note that the results from negative class are improved as the number of samples increase. For the class "P", the recall and precision are directly related. That is: if the precision grows, the recall decreases and vice-versa. We believe that the best results for the positive class can be achieved in the proportion of 1 to 1 (1 positive pair for each negative one) and 1 to 2 (1 positive pair for each 2 negative ones). This means "1 to 1" for a more balanced precision and recall; and "1 to 2" for a more precise classification.

7. Conclusion

In this paper, we aimed to describe the generation of a model for solving coreferences in Portuguese, focusing on categories of named entities and specific features. Our balancing levels "1 to 1" (1 positive for each negative one) and "1 to 2" (1 positive for each 2 negative ones) are promising in comparison to our baseline model, in which the results for the class "P" is 74.3% of precision, 30.7% of recall and 43.4% of f-measure. In these two balancing levels, we obtained 79.5% of precision, 56.0% of recall and 65.6% of f-measure in "1 to 1" proportion and 90.6% of precision, 40.1% of recall and 55.6% of f-measure in "1 to 2" proportion. The global result is also good, if compared with Fonseca et al.'s work. They focused only in pairs containing proper nouns in both noun phrases, achieving a precision of 76.8%, a recall of 88.9% and an f-measure of 82.4%. In our current work, we increase this scope, solving co-reference for all noun phrase pairs.

In order to validate our experiments, we ran each of the five experiments with different under-sampling size 100 times, calculating an average from precision, recall and f-measure. In our experiments, we proved that the under-sampling technique may help coreference resolution, improving the results. As future work, we want to test semantic resources, such as Onto-PT (Oliveira et al., 2014, in order to introduce a richer semantic discourse at computational level.

Acknowledgements

The authors acknowledge the financial support of CNPq, CAPES and Fapergs.

References

Abreu, S. C., Bonamigo, T. L., and Vieira, R. (2013) A review on Relation Extraction with an eye on Portuguese, Pages 1-19 In:

Journal of the Brazilian Computer Society. Available at: http://link.springer.com/article/10.1007%2Fs13173-013-0116-8#page-1

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. In Machine Learning: ECML 2004 (pp. 39-50). Springer Berlin Heidelberg.

Bick, E., (2000) The Parsing System "Palavras" - automatic grammatical analysis of portuguese in a constraint grammar framework, Tese de Doutorado, Department of Linguistics, University of Århus, DK.

Cimiano, P., (2006). Ontology learning and Population from text. Algorithms, Evaluation and Applications. University of Karlsruhe Inst. AIFB, Germany. ISBN 0387306323

Boukckaert, R., Frank, E., Hall, M., Kirkby, K., Reutemann, P., Seewald, A. and Scuse, D., (2013) Weka Manual for version 3.6.9, The University of Waikato. Available at: http://www.cs.waikato.ac.nz/ml/weka/

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. In Journal of Artificial Inteligence Research 16:321-357

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter, 6(1), 1-6.

Cieslak, D. A., & Chawla, N. V. (2008). Learning decision trees for unbalanced data. In Machine Learning and Knowledge Discovery in Databases (pp. 241-256). Springer Berlin Heidelberg.

Collovini, S., Carbonel, T., Fuchs, J., Coelho, J., Rino, L. and Vieira, R. (2007) Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática. In: V Workshop em Tecnologia da Informação e da Linguagem Humana – TIL. Proceedings of XXVII Congresso da SBC, Rio de Janeiro.

Coreixas T., Resolução De Correferência E Categorias De Entidades Nomeadas, Dissertação De Mestrado, Pontifícia Universidade Católica Do Rio Grande Do Sul, 2010. Available at: http://repositorio.pucrs.br/dspace/handle/10923/1567

CoNLL2011, Conference on computational natural language learning, Available at: http://conll.cemantix.org/2011/. Access on: 05/08/2012.

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, *20*(1), 18-36. Available at http://150.214.190.154/docencia/doctoM6/Estrabrooks.pdf

Fonseca E. B., Vieira R., Vanin A. (2013), Resolução de Correferência em Língua Portuguesa: Pessoa, Local e Organização. In X National Meeting on Artificial and Computational Intelligence.

Freitas, C., Mota, C., Santos, D., Oliveira, H. and Carvalho, P. (2010) Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese, Linguateca, FCCN.

Gabbard, R., Freedman, M. and Weischedel, R. M. (2011) Coreference for Learning to Extract Relations: Yes, Virginia, Coreference Matters. In: Proceedings 49th Annual Meeting of the Association for Computational Linguistics: shortpapers, pages 288–293, Portland, Oregon.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D. (2011) Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task, Conference on computational natural language learning.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics, 39(4), 885-916.

Ling, C., and Li, C. (1998). Data Mining for Direct Marketing Problems and Solutions. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98) New York, NY. AAAI Press.

Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. In Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning Las Vegas, Nevada.

Miller, G. (1995) WordNet: A Lexical Database for English. In: Communications of the ACM Vol. 38, No. 11: 39-41. Available at: http://dl.acm.org/citation.cfm?id=219748

Oliveira, H. G., & Gomes, P. (2014). ECO and Onto. PT: a flexible approach for creating a Portuguese wordnet automatically. Language Resources and Evaluation, 48(2), 373-393. Available at http://link.springer.com/article/10.1007/s10579-013-9249-9/fulltext.html

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R. and Xue, N. (2011) CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes, CoNLL Shared Task.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47. Available at: http://arxiv.org/pdf/cs.ir/0110053.pdf