Learning Discourse-new References in Portuguese Texts

Sandra Collovini and Renata Vieira

Universidade do Vale do Rio dos Sinos
CP 275, CEP 93022-000, São Leopoldo, RS, Brazil
sandrac@exatas.unisinos.br,renatav@unisinos.br

Abstract. This work presents the evaluation of a discourse status classifier for the Portuguese language. It considers two distinguished classes of discourse novelty: *Brand-new* and *New* references. An evaluation of the relevant features according to different linguistic levels are presented in detail.

1 Introduction

The identification of discourse status has been recognized as a relevant task in natural language understanding. Many systems have been proposed to classify referring expressions [2, 14, 12, 5, 13, 8] in order to recognize if they are new or old information. This comes along with the problem of anaphora resolution, it is usually useful establish the relations of old expressions with their antecedents. It is important, for instance, to identify antecedents for pronouns (it, he, she) to interpret the meaning of the discourse. Our work focuses on definite descriptions (DDs), those referring to expressions with a definite article (such as *the boy, the girl*) , because they are numerous in texts and are the main source of ambiguity regarding novelty, as opposed to other expressions. Pronouns, for instance are mainly old and indefinite descriptions are mainly new.

Whereas most of the literature in this area refers to the English language, we built and evaluated a system to classify discourse status in Portuguese texts. Besides proposing and evaluating such a system for a new language, this work is original by considering two different classes of discourse-new DDs. At first, we classified *Brand-new* definite descriptions. However, as the distinction between *Brand-new* and *Anchored-new* DDs is remarkably difficult [9], a second study was made considering the more general class *New*, which includes both *Brand-new* and *Anchored-new*. Also original in this study is that the relevance of the features used for learning the classifier is analyzed considering different levels of linguistic knowledge.

This paper is organized as follows. In Section 2 we discuss the related work. Classes of discourse status are defined and exemplified in Section 3. In Section 4, a corpus study and the features used to build our classifier are presented. In Section 5 we discuss the resulting decision trees and the relevance of the features is discussed in Section 6. In Section 7, this work also shows an evaluation of the resulting system on completely unseen data. In Section 8 we present our final remarks.

2 Related Work

There are in the literature several proposals of referring expressions classifiers. In [2] a classifier for DDs was developed. The authors define new DDs as independent existential expressions, understood by the readers isolately, without needing a context. This system conjugates 9 syntactic heuristics (restrictive pre-modifiers and post-modifiers, relative clauses, adjective constructions etc.) and other heuristics like DDs that occur in the first sentence of a text. As a result, they achieved 78% of recall, 87% of precision and 82% of F-measure for the classification of independent existing DDs. In [14] a heuristic based discourse-new DD classification system was developed, reaching 69% of recall, 72% of precision and 70% of F-measure, on the basis of 9 such features.

In [13] a classifier for discourse-new DDs and unique expressions was presented, where discourse new DDs were defined as the first mention of an entity in the discourse and unique expressions were said to specify their referent totally, and for this reason are understood without any context. The author took into consideration 32 features (syntactic, contextual and definite probability) including data from the web. The reported result was 82.3% of recall, 84.8% of precision and 83.5% of F-measure in discourse-new DDs classification and 68.8% of recall, 85.2% of precision and 76.1% of F-measure were reported for unique expressions classification. In [8] a group of common features in these previous work for another discourse-new DDs classifier (9 features) was reviewed and applied. The classifier resulted in 95.1% of recall, 85.8% of precision and 90.2% of F-measure.

All works cited above refer to the English language. Some other languages are also studied but not so extensively [1, 4, 5]. There are some corpora studies about coreference for the Portuguese language [11], but to the best of our knowledge there is no implemented DD resolution or classification system for Portuguese, so far. In addition to that, another difference of our work is that we give a detailed analysis of the features that were actually relevant to the classification, whereas in these previous work there is usually none. An exception is [7], which examines anaphoricity information to improve a learning-based coreference system and presents a list of the most informative features.

This work is the first one to present a classifier for DDs in Portuguese language. Based on a corpus study, DDs were analyzed and a set of features was organized in 3 groups considering three distinct linguistic levels. Features specifically related to the noun phrase structure constitute the first group, features which consider the sentence structure are in the second one, and the third group is based on information about the previous sentences. In the next section, we present the classes in detail.

3 Classes Description

The classes of DDs considered in this work are mainly based on [10], but they are also related to many of the studies discussed in Section 2. In the examples below, DDs are presented in boldface and their antecedents are underlined.

New DDs: these are definite referring expressions that introduce new entities into the discourse. In this work we consider two types of New DDs, *Brand-new* and *Anchored-new*.

- **Brand-New DDs** (discourse-new or non-anaphoric): introduce entities which are new in the discourse:
  A Folha de São Paulo *apresentou as listas apreendidas na operação contra o crime organizado.* ***[The Folha de São Paulo*** *presented the lists arrested in the operation against the organized crime.]*
- **Anchored-new DDs** (associative anaphors or bridging): refer to entities that have a semantic connection with an antecedent expression, which is necessary to their interpretation:
  A Folha de São Paulo apresentou as listas apreendidas na operação contra o crime organizado. O jornal tentou ouvir ***o delegado encarregado****. [The Folha de São Paulo presented the lists arrested in the operation against the organized crime. The newspaper tried to listen to* ***the police chief in charge.]***

Old DDs: refer to entities mentioned in the previous discourse. Old DDs can be *Plain-old* and *Related-old*.

- **Plain-old DDs** (direct anaphors): have an identity relation with their antecedents and share with them the same head-noun:
  *... as listas apreendidas na operação contra o crime organizado. Alguns delegados também são citados n****as listas****. ...* ***[the lists arrested in the operation against the organized crime****. Some police chiefs are also mentioned in the lists.]*
- **Related-old DDs** (indirect anaphors): have an identity relation with their antecedents however they present a distinguished head-noun:
  A Folha de São Paulo apresentou as listas apreendidas ... ***O jornal*** *tentou ouvir .... [The Folha de São Paulo presented the lists arrested ...* ***The newspaper*** *tried ...]*

4 Corpus study

Our work was based on two corpora. Corpus 1 was formed by 24 newspaper articles from Folha de São Paulo, written in Brazilian Portuguese, corresponding to part of the NILC[1] corpus. Out of 2319 noun phrases (NPs) we identified 1331 DDs. Corpus 1 was used for the learning phase. Corpus 2 was composed by 4 texts from the Public newspaper, written in European Portuguese from CETEMPublico[2] corpus. Out of 770 noun phrases we identified 482 DDs. Corpus 2 was used for the final evaluation.

The corpora were automatically annotated with syntactic information using the parser PALAVRAS[3] [3] to Portuguese. They were also manually annotated with coreference using MMAX [6]. The first annotation task was to distinguish *New* and *Old* DDs. The second task was pointing to the antecedent for the old cases. Corpus 1 was

[1] http://www.nilc.icmc.usp.br
[2] http://www.linguateca.pt/CETEMPublico
[3] http://visl.sdu.dk/visl/pt/parsing/automatic

annotated by three annotators. The agreement for the first task was close to 90.0%. Corpus 2 was annotated by four annotators. For the first task, the agreement resulted in 94.7% among the four annotators, for all other cases there was agreement among three annotators. This two-fold distinction is much easier than for the four classes, which explains why agreement was high whereas other work usually report much less than that. For the second task, antecedents annotation for those classified as old, four annotators agreed in 73.9% of the cases, in other 6.3% of the cases there was agreement among three annotators, in 0.84% only two annotators agreed, and complete disagreement was verified for the remaining 18.9%. The results are shown in Table 1.

Table 1: Manual Annotation

Corpus	**New DDs** (%)	**Old DDs** (%)	**Total** (%)
1	816 (61.3%)	515 (38.7%)	1331 (100%)
2	308 (63.9%)	174 (36.1%)	482 (100%)

The corpus was further analyzed, dividing *New* and *Old* DDs in their subclasses as presented in Section 1 (see Table 2). The usual large quantity of *Brand-new* DDs was confirmed. In Corpus 1, 52.3% were *Brand-new* and in Corpus 2 this number was even higher, 59.5%. DDs of Corpus 1 were studied against the features described in previous work, as presented in Section 3. A total of 16 features were identified in three groups of features according to different levels of linguistic knowlege. Group G1 considers information about the noun phrase alone, G2 considers information about the sentence in which the DD appears, G3 takes into account information about the previous text detailed in Table 3. Examples from the corpus illustrating each of the features are presented.

– PP: *Os membros da classe jurídica. [The members of the juridical class.]*
– APP: *O Prefeito de Gravataí, Daniel Luiz Bordignom. [The Gravataí major, Daniel Luiz Bordignom.]*
– PN_APP: *O delegado Elson Campelo. [The Police Chief Elson Campelo.]*
– REL_CL: *O texto que deve ser assinado pelos jornalistas. [The text that must be signed by jornalists.]*
– CPN_HEAD: *O Othon Palace Hotel. [The Othon Palace Hotel.]*
– AP: *As conversas mais antigas. [The older conversations.]*
– ADJ_PRE: *O primeiro grau. [The first degree.]*
– NUM_PRE: *Os 65 anos. [The 65 years.]*
– NUM: *Os anos 60. [The sixties (decade).]*
– PRON_DET: *Os nossos arqueólogos. [The (our) archaeologists.]*
– SUP_PRE: *Os melhores alunos. [The best students.]*
– SUP: *O Christofle líquido é o melhor. [The Liquid Christofle is the best.]*
– SIZE: *O quilômetro 430 da rodovia Assis Chateau Briand. [The 430 Km from Assis Chateau Briand road.]*
– COP: *O coreano seria a língua dos anjos. [(The) Koren would be the angels tongue.]*

These features were used for decision trees learning on the basis of examples from Corpus 1. After the learning process, the best resulting trees were implemented and further tested on unseen data (Corpus 2).

Table 2: New and Old subclasses

Corpus	**New DDs**		**Old DDs**	
	B-new (%)	A-new (%)	P-old (%)	R-old (%)
1	696 (52.3%)	120 (9.0%)	364 (27.35%)	151 (11.3%)
2	287 (59.5%)	21 (4.4%)	159 (33.0%)	15 (3.1%)

Table 3: Groups of Features

Groups	**Feature**	**Description**
G1	PP	Prepositional phrase.
	APP	Apposition.
	PN_APP	Appositive proper name with no explicit mark.
	REL_CL	Relative clause.
	CPN_H	When the head is a compound proper name.
	AP	Adjectival phrases.
	ADJ_PRE	Adjective preceding the head.
	NUM_PRE	Number before the head.
	NUM	Number after the head.
	PRON_DET	Other determinant besides the definite article.
	SUP_PRE	Superlative premodifier.
	SUP	Superlative alone.
	SIZE	Containing five terms or more.
G2	COP	DDs in a copular construction.
	S1	DDs that occur in the first sentence of the text.
G3	NO_ANT	DDs head is a word that does not occur previously in the text.

In [8] a set of 9 features from 6 groups (anaphora, predicative NPs, proper names, functionality, establishing relative, text position) was proposed. Our study takes 3 groups of features which are different from those presented in [8], but the features themselves are similar. They consider proper name, apposition, prepositional phrase, relative clause, superlative, copular construction, position in text, and anaphora. Our choice of 3 groups was motivated by the analysis of the NP alone, the NP plus sentence structure and position, and the NP, sentence plus previous text.

5 Decision Trees Learning

The learning algorithm used was Weka[4] *j48*, with 10 fold cross-validation. We tested different combinations of the 3 group of features for the decision trees generation: G1, G12 (=G1+G2) and G123 (=G1+G2+G3). Group G1 considers the noun phrase alone, G12 considers the noun phrase features and also information about the sentence, G123 will take into account noun phrase and sentence information but also the existence of a noun phrase with the same head as the DD in the previous text.

The first classification experiment considered the classes *Brand-new* (expressions that do not have an antecedent) and *Other* (expressions that have an antecedent). The results are presented in Table 4 and the features considered for the resulting trees in Table 5, in order of appearance in the trees. G123 presented the best results of precision, recall and F-measure for the *Brand-new* class, and the higher number of correctly classified occurrences in general. G1 alone, however, results in precision as high as other groups. It is in recall that G123 shows improvements when compared to the others. The number of features went down to 4 in G123.

Table 4: Brand-new classification
Correct(C); Precision (P); Recall (R); F-measure (F)

Classes	**G1**				**G12**				**G123**			
	C	**P**	**R**	**F**	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**
B-new	63%	65%	55%	60%	64%	66%	57%	61%	70%	65%	88%	75%
Other		61%	70%	65%		62%	71%	60%		82%	53%	64%

Table 5: Features for classifying Brand-new DDs

Relevant features	
G1	SIZE, AP, CPN_H, ADJ_PRE,NUM_PRE, PN_APP
G12	S1, SIZE, AP, ADJ_PRE,CPN_H, PN_APP
G123	S1, NO_ANT, SUP_PRE, NUM

Table 6: New classification
Correct(C); Precision (P); Recall (R); F-measure (F)

Classes	**G1**				**G12**				**G123**			
	C	**P**	**R**	**F**	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**
New	61%	71%	58%	64%	61%	71%	61%	66%	77%	76%	89%	82%
Other		53%	66%	59%		55%	66%	60%		81%	60%	69%

[4] http://www.cs.waikato.ac.nz/ml/weka

Table 7: Features for classifying New DDs

Relevant features	
G1	SIZE, NUM, PN_APP, AP, CPN_H,ADJ_PRE, PP, NUM_PRE
G12	S1, SIZE, PN_APP, NUM, AP, CPN_H,ADJ_PRE, PP, COP
G123	NO_ANT, NUM, S1, SUP_PRE

The second classification considered the classes *New* (including both *Brand-new* and *Anchored-new*) and *Other*, corresponding to *Old*. The results are presented in Table 6. Results were all higher than for *Brand-new*. G123 shows higher precision and a much higher recall than the other groups. The number of resulting attributes was again 4 in G123 (see Table 7).

6 Feature Analysis

Tables 5 and 7, in the previous section, show the features included in the generated decision trees. The larger number of attributes in a tree was 8 and 9, for G1 and G12. When NO_ANT was considered, this number went down to 4. Features APP, REL_CL, PRON_DET, SUP were never included in the resulting trees. The attributes were evaluated separately to verify which of them contributed individually and strongly for the classification.

The prominent features for *Brand-new* DD classification of each group are displayed in Table 8. In G1, SIZE was a feature that, alone, was able to reach 44% F-measure, with 67% precision. S1 in G2, although has shown 100% precision, is of limited recall, since it only applies to the first sentence of each text. In G3, NO_ANT had 73% F-measure and 64% precision. The SIZE feature is an original attribute that is simple to be verified and has presented a significant precision result if compared to the entire group G1 and also with higher precision than NO_ANT of G3. For these reasons, we analyzed decision trees generated on the basis of G1 but without the SIZE feature (G1 without SIZE), in Table 9. We noticed that the feature SIZE replaces other features commonly present in related work (prepositional phrases, relative clauses) in a satisfactory way and presents increases in the number of correctly classified descriptions and in precision in the classification of *Brand-new* DDs. When SIZE is not considered, the resulting tree includes PP, ADJ_PRE, REL_CL, which didn't appear before.

Table 8: Feature analysis
Precision (P); Recall (R); F-measure (F)

Feature Alone	**P**	**R**	**F**
SIZE (G1)	67%	33%	44%
S1 (G2)	100%	6%	11%
NO_ANT (G3)	64%	86%	73%

Table 9: Feature SIZE
Correct C); Precision (P); Recall (R); F-measure (F)

Features	**C**	**P**	**R**	**F**
G1	63%	65%	55%	60%
G1 without SIZE	62%	63%	58%	61%

In the *New* DD classification, the only feature that presented a distinction when applied alone was NO_ANT with 76% of correct classification, precision of 76% and recall of 86%. Other features alone were not able to distinguish the examples. When the previous text is considered as a feature, the features related to the noun phrase structure seem to loose their importance for the task.

7 Evaluation on unseen data

The decision trees learned in the experiments shown in the last section were applied to completely unseen data - Corpus 2. So we could also check the adequacy of the learned trees for this variant of Portuguese. The results are presented below.

The results of the *Brand-new* classifier applied to Corpus 2 can be seen in Table 10. We adopted as baseline (B) an algorithm that classifies all definite descriptions as *Brand-new*. As before, group G123 showed the best results. The difference from G123 to G1 and G12 was significant (99.5%). We verified significant gains in precision (from 60% to 86%) and F-measure (from 75% to 80%) considering the given baseline. Note that for the *Other* class, F-measure was never lower than 66%. G1 alone shows improvements in precision compared to the baseline (from 60% to 80%).

Table 10: Brand-new Classification
Correct(C); Precision (P); Recall (R); F-measure (F)

Classes	**B**				**G1**				**G12**				**G123**			
	C	**P**	**R**	**F**	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**
B-new	59%	60%	100%	75%	68%	80%	62%	70%	69%	80%	64%	71%	78%	86%	76%	80%
Other		0	0	0		58%	77%	66%		59%	76%	66%		70%	82%	75%

For the class *New*, the results of Group G123 are significantly higher than the others (99.5%), 83% of precision and 85% of F-measure, against a baseline of 64%, and 78% (see Table 11). Again, group G1 presents improvements in comparison to the baseline (from 64% to 80%).

The results reported are even better than the ones shown for the learning phase, this is probably related to the higher number of *Brand-new* and *New* DDs in the European Portuguese Corpus (Table 2). Features related to the noun phrase structure have been used in many of the previous work, and we can see here that they alone can indicate, with considerable precision, the novelty level of DDs.

Table 11: New Classification
Correct (C); Precision (P); Recall (R); F-measure (F)

Classes	**B**				**G1**				**G12**				**G123**			
	C	**P**	**R**	**F**	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**
New	64%	64%	100%	78%	65%	80%	61%	70%	67%	79%	66%	72%	81%	83%	88%	85%
Other		0	0	0		52%	73%	61%		54%	70%	61%		76%	69%	72%

8 Final Remarks

This work presented the evaluation of a classification system of *Brand-new* and *New* DDs for Portuguese. The evaluation was carried out on completely unseen data. The results were stable. Classifying *New* DDs seems to be easier than classifying *Brand-new* DDs, as we can see higher F-measure values for this class (although this was clearer in the first experiments with corpus 1). In the classification of *Brand-new* DDs, Group G123 has shown a F-measure of 80%. Group G1 has shown a precision of 80%. Group G12 doesn't show much improvement due to the limited number of cases in copular constructions and in first sentences. In the classification of *New* DDs, the attributes in G123 showed a F-measure of 85%. In G1, the precision is 80%, near to 83% seen in G123.

We were interested in the contribution of the noun phrase alone for the classification (G1), and we found that it was indeed enough for achieving high precision. These findings might have interesting consequences for other tasks, such as summarization. In an extracted summary, for instance, DDs can be analyzed solely according to their intrinsic structure, to verify if they are new in the discourse. In these cases they would not bring problems of coherence to the summary due to the lack of an antecedent.

A detailed evaluation of the features was made. We found that the feature SIZE alone presented a better precision than other features in Group 1 altogether (67%). This feature seems to replace well several complex syntactic features often used in other systems, such as relative clauses and prepositional phrases. It is a simple feature that has not been mentioned in previous work so far. The feature NO_ANT (G3) was rather relevant in both classifications, confirming the findings of [7] for English. In fact, when classifying *New* DDs it is the only salient feature. Also, this feature minimizes the importance of other features. Indeed, looking for the presence of an identical antecedent seems to do alone most of the job.

We acknowledge that related work deal with different kinds of NPs, different features, languages and data. This of course makes the comparison difficult. However, we can see that, in general, the results of the proposed system are not far from the state of the art in the area as reported by previous work (Table 12). From a initial set of 16 features our classifier achieved best measures on the basis of 4 of them. As future work, we intend to carry out an investigation into other romance languages and other classes (*Plain-old, Related-old, Anchored-new*).

Table 12: Related work

Related Work	**P**	**R**	**F**	**#Features**
[2] - Independent existential DDs	87%	78%	82%	10
[14] - Discourse new DDs	72%	69%	70%	9
[13] - Discourse new DDs	85%	82%	83%	32
[8] - Discourse new DDs	95%	86%	90%	9
We - Brand-new DDs	86%	76%	80%	16/4
We - New DDs	83%	88%	85%	16/4

References

1. C. Aone and S. Bennett. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the ACL*, pages 122–129, Cambridge, Massachusetts, USA, 1995.
2. D. L. Bean and E. Riloff. Corpus-based indentification of non-anaphoric noun phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 373–380, College Park, Maryland, USA, 1999.
3. E. Bick. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Protuguese in a Constraint Grammar Framework*. PhD thesis, Arhus University, Arhus, 2000.
4. R. M. Guillena, M. Palomar, and A. Ferrández. Processing of spanish definite descriptions. In *Proceedings of the Mexican International Conference on Artificial Intelligence*, pages 526–537. Springer-Verlag, 2000.
5. C. Müller, S. Rapp, and M. Strube. Applying co-training to reference resolution. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 352–359, Philadelphia, PA, 2002.
6. C. Müller and M. Strube. Mmax: A tool for the annotation of multi-modal corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 45–50, Seattle, Washington, 2001.
7. V. Ng. Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 151–158, Barcelona, Spain, 2004.
8. M. Poesio, M. Alexandrov-Ksbadjov, R. Vieira, R. Goulart, and O. Uryupina. Does discourse-new detection help definite description resolution? In *Proceedings of the 6th International Workshop on Computational Semantics*, pages 236–246, Tiburg, 2005.
9. M. Poesio and R. Vieira. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, 1998.
10. E. F. Prince. Toaward taxonomy of given-new information. In *P. Cole, editor Radical Gramatics*, pages 223–256, New York, 1981. Academic Press.
11. S. Salmon-Alt and R. Viera. Nominal expressions in multilingual corpora: Definites and demonstratives. In *Proceedings of the LREC*, pages 1627–1634, Las Palmas, 2002.
12. W. M. Soon, H. T. Ng, and D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases. In *Computational Linguistics*, volume 27, pages 521–544, 2001.
13. O. Uryupina. High-precision identification of discourse new and unique noun phrases. In *Proceedings of the 41st Annual Meeting on ACL*, pages 80–86, Sapporo, Japan, 2003.
14. R. Vieira and M. Poesio. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):525–579, 2000.