Paramopama: a Brazilian-Portuguese Corpus for Named Entity Recognition

Carlos Mendonça Júnior
Departamento de Computação
Universidade Federal de Sergipe
São Cristovão, Sergipe
Email: juniorcaemj@gmail.com

Hendrik Macedo
Departamento de Computação
Universidade Federal de Sergipe
São Cristovão, Sergipe
Email: hendrik@ufs.br

Thiago Bispo
Departamento de Computação
Universidade Federal de Sergipe
São Cristovão, Sergipe
Email: thiagodiasbispo@gmail.com

Flávio Santos
Departamento de Computação
Universidade Federal de Sergipe
São Cristovão, Sergipe
Email: flavioarthurufs@gmail.com

Nayara Silva
Departamento de Computação
Universidade Federal de Sergipe
São Cristovão, Sergipe
Email: nayararosy.fs@gmail.com

Luciano Barbosa
IBM Research
IBM Brasil
Rio de Janeiro, Rio de Janeiro
Email: lucianobarbosa@gmail.com

***Abstract*—Named Entity Recognition (NER) is one of the most important Natural Language Processing (NLP) tasks. NER concerns the ability to automatically identify words or pieces of text that refers to some desired target entity, such as name of places, people, organizations and time units or event dates. NER is central to Information Extraction (IE) applications. Despite that, there are few NE-tagged corpora for Brazilian Portuguese (PtBR). This paper proposes a new tagged Brazilian Portuguese corpus for Named Entity Recognition, which we call Paramopama. We evaluate the quality of Paramopama corpus by measuring Precision, Recall and F-measure of an NER classifier trained from this corpus. In our experiments, an NER classifier built from Panorama corpus has achieved better F-measure results compared to NER classifiers built from other well-known PtBR NE-tagged corpora.**

I. Introduction

Named Entities (NE) [6] are basically anything that can be referred to with a proper name belonging to predefined categories such as the names of persons, organizations, locations, expressions of time, quantities, monetary values, percentages, among others. NER plays a central role for extracting relevant information from a text. In [12], for instance, NER recognition is used to provide the recognition of biomedical data such as proteins, DNA and RNA data and type of cells from medical articles.

As a typical automatic classification task, accurate NE detection requires a good training set. In a training set, all the terms that represent desired entities should be properly labeled. Usually, the labeling process is a hard-working and error-prone task. Luckily, there is a large amount of NE *corpora* available meeting different languages: English[14], [15], [3], [8], Spanish [13], German [14], Dutch [13], Chinese [8], and Arabic [8].

For Portuguese, however, there is a lack of corpora with Named Entity tagging - from now on, we will refer to it as *PtBR NE-tagged corpora*. The most relevant corpora, certainly the most cited one, are the HAREM corpora [11], [5]. Combined, they add up to 10,000 sentences with 200,000 words tagged in 9 different categories. The best HAREM results achieved an F-measure of 70%. For comparison, English corpora have currently reached an F-measure of 88% [14]. In this work, we will use only 4 classes of the HAREM corpus (Person, Location, Organization and Time) for better comparison among all corpora.

Despite of its small size, HAREM is said to be a gold-collection, since it has been manually tagged from scratch by humans. There are many solutions that describe different methods to automatically tag a large amount of text. The corpus resulting of such approaching is said to be *silver-standard*. [10] proposes a method to create NE-tagged an corpus by exploiting the text and structure of Wikipedia, called WikiNER. Basically, this method classifies Wikipedia articles in named entities, and then transforms the links between articles into NE annotations considering the wikilinks at the pages. In addition, inter-languages links were used to generate nine different language corpora, including PtBR. A limitation of this proposal is that it does not take into account sentence context, causing inconsistencies in class labeling for some sentences, such as confusion between the entities *Localization* and *Organization* or *Person* and *Localization*. For instance, in the sentence: "Em 1964, O Reino Unido concede a independência a Rodésia do Norte, com o nome de Zâmbia.", the words "Reino Unido" and "Rodésia do Norte" were labelled as Location in the WikiNER corpus. But in this context, they should be labelled as Organization, because they refer to a nation state. On the other hand, the word "Zambia" was properly labeled as Location, because in this context refers to a country name. Another limitation, is the insufficient number of examples for the Organization entity, resulting in a poor performance to classify words of this class.

This paper extends the PtBR version of WikiNER corpus, revising incorrect assigned tags in order to improve corpus quality. We also extend the corpus size and provide proper evaluation. We name this new corpus as Paramopama.

The rest of the paper is organized as follows. Section 2 presents the method for corpus construction as well as the training for the learning model. Experiments conducted to proper evaluate the corpus and to compare with others of the state of the art are presented in Section 3. Results are also discussed. Finally, we conclude the work in Section 4.

II. Method

A. Corpus generation

We implemented a two-step process to build the corpus: improving and expanding, as presented in Figure 1.

- **Improving**. In this phase, we use an NE classifier trained with HAREM corpus, to automatically label the WikiNER corpus. The result of this auto-labeling process was then revised, comparing the two labels in each word of each sentence and manually removing the wrong named entity. Furthermore, we have also added the tag TIME to the corpus to identify the time expressions. At end of this process, we produced, as a whole, a set of 10,000 sentences of the WikiNER corpus which had their tags corrected and revised.
- **Expanding**. In this phase, we used the set produced in the improving phase to train an NER classifier and label around 2,500 newer sentences which were obtained from news websites of various different domains (e.g. economics, politics, sports, technology). In particular, we have biased the crawling process to the presence of Organization entity instances. This was important in order to improve classifier's learning ability for this entity, thereby reducing the number of Organization false negatives and decreasing the number of false positives of the other entities. Next, the resulting NE classifier was used to label the newer sentences. Finally, the words labelling was reviewed(as in the improving process) and the sentences were added to the final corpus.

The result of this twofold procedure is the fully NE-tagged corpus, Paramopama, composed by revised sentences of WikiNER corpus and a set of news sentences, tagged and reviewed. Tables I and II compares the size of Paramopama corpus to three other mainstream PtBR NE-tagged corpora in regards to the number of sentences and the number of tagged entities, respectively.

Paramopama corpus was tagged with *Person*, *Location*, *Organization* and *Time* tags. The other words were tagged as *Other*. Figure 2 shows excerpts from Paramopama.

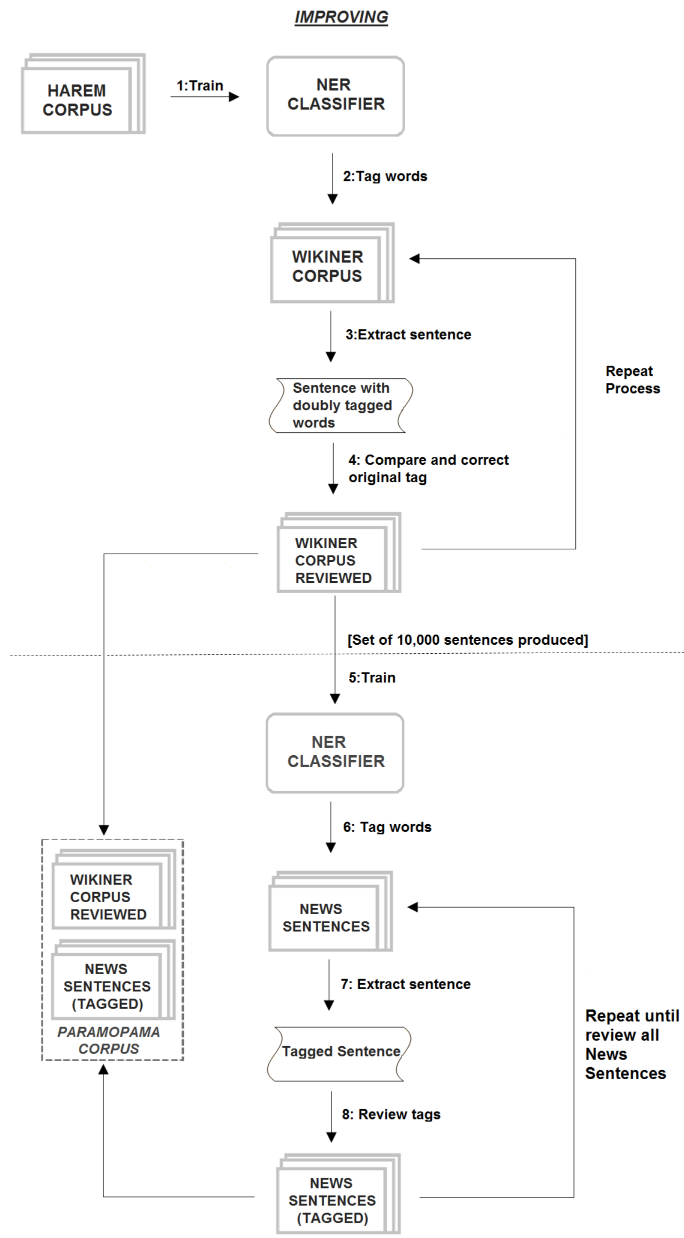

Fig. 1. Improving and Expanding phases of corpus generation.

TABLE I. Size of PtBR NE-tagged corpora

Corpus	Sentences	Tokens
Paramopama	12.500	310.000
Primeiro HAREM	5.000	80.000
Segundo HAREM	3.500	100.000
WikiNER	125.821	2.830.000

B. NER Classifier

Choosing a machine learning tool is an important decision for getting good recognition rates. We chose Stanford NER[1] for the following reasons. It is a sequence classifier which implements the Conditional Random Fields (CRF) of arbitrary order [7]. The rationale for the choice for CRF is that CRF has the ability to mitigate the strong independence assumption

TABLE II. SIZE OF PTBR NE-TAGGED CORPORA (NUMBER OF TAGGED ENTITIES)

Corpus	PER	LOC	ORG	TIME
Paramopama	7.327	17.461	7.154	10.827
First HAREM	2.244	2.169	2.039	914
Second HAREM	4.018	1.930	1.607	3.291
WikiNER	94.394	199.806	36.086	-

of Hidden Markov Models (HMM) and Stochastic Grammars, for instance. It also avoids an important issue of Maximum Entropy Markov Models (MEMM): a biased behavior in states that are preceded by few states [7]. Furthermore, previous work shows good results of CRF to NER [9], [12] on English corpora.

In addition to the correctly tagged corpus and the choice of a proper learning model, defining a relevant set of features is crucial. Features should be plausible predictors of the class tag and should be easily and reliably extracted from the corpus. Some of these features are related to the machine learning model such as the sliding window of Viterbi algorithm, in which a forward positioned word may influence the tag of a previous one. Other features are related to the token to be tagged and to the surrounding text. For instance, words with mixed case or punctuation, like *eBay* and *Yahoo!*, are strong indicators that the word should be tagged as Organization. A date, on the other hand, can be easier identified with the use of regular expressions, so they are tagged in the Time class. Features that define gazetteers of cities, countries or monuments can be used to tag words in the Location class. Similarly, gazetteers of person names can be used to tag words in the Person class.

We have defined several features in our Named Entity Classifier. Most of them refers to lexical items relative to word shape feature such as word case, presence of digit, punctuation or hyphen, lemmas, stems and word letters n-grams, very important for training the classifier in detection of particular classes. For instance, person names have its first letter in capital, words with mixed case, like eBay, probably must be classified as Organization. We have also set gazetteers for person names, cities, countries and organizations collected from websites and databases available on the Internet. For example, we considered the 2014 Forbes list of biggest organizations in the world. These gazetteers are essential for detection of all classes, especially person names. Other features refer to the context of the word to be tagged, such as the previous word, the following word, previous word's tag, among others. Also has been set the class feature, relative to the frequency of each class in the corpus. The CRF order is also an important feature because it defines the context window where all features will be evaluated. Table III lists some of the features we have defined to represent the instances in the training set.

III. EVALUATION AND RESULTS

The goal of our experimental evaluation is to assess the Paramopama corpus, comparing the results with other three PtBR NE-tagged corpora and combinations of them: First HAREM, Second HAREM, First + Second HAREM (Full HAREM), WikiNER, Paramopama and Paramopama +

Sentence 1		Sentence 2	
A	O	Águeda	LOCAL
bandeira	O	dista	O
nacional	O	de	O
,	O	Lisboa	LOCAL
adotada	O	240	O
em	TEMPO	km	O
1947	TEMPO	,	O
,	O	do	O
é	O	Porto	LOCAL
baseada	O	72	O
na	O	km	O
bandeira	O	e	O
do	O	de	O
Congresso	ORGANIZACAO	Aveiro	LOCAL
Nacional	ORGANIZACAO	cerca	O
Indiano	ORGANIZACAO	de	O
,	O	20	O
desenhada	O	km	O
por	O	.	O
Pingali	PESSOA		
Venkayya	PESSOA		
.	O		

Fig. 2. An excerpt from Paramopama

HAREM. Precision, Recall and F-Measure (F1-Score) are the evaluation measures.

Two different test sets were considered. The first test set consisted of the last 10% of Second HAREM corpus, accounting for 383 sentences. The second test set consisted of 5,855 sentences from WikiNER corpus that have been improved as the result of the Improving phase.

In both test sets, the named entities used in the evaluation are PERSON, LOCATION, ORGANIZATION and TIME. Actually, WikiNER corpus is not evaluated in regards to TIME entity since this tag was not present in the original corpus.

Tables IV, V, VI and VII show the values of Precision, Recall and F-measure for the first test set (Set 1) and Tables IX, X, XI and XII show the values of Precision, Recall and F1-Score for the second test set (Set 2), for the named entities PERSON, LOCATION, ORGANIZATION and TIME, respectively. Tables VIII and XIII show the overall mean result for Set 1 and Set 2, respectively.

The results show that, in general, *Paramopama*'s NER classifier performs better than both HAREM and WikiNER. The overall score for test set 1, presented in Table VIII, indicates an F-measure around 80% for *Paramopama*'s whereas HAREM has achieved at most around 77% and WikiNER with less than 60%. Similar comparative performance has been achieved for the second test set, in Table XIII. A combination of Paramopama with HAREM has achieved the best results of all, with an outstanding performance for the TIME class with an average F-Measure of 92% (Tables VII and XII).

Precision values achieved by *Paramopama*'s NER classifier (Tables IV, V, VI and VII) were mostly above 80 %, which means a low false positive rate in a text classification task, especially if we compare with WikiNER's, which achieved average Precision of 60%.

Another important result depicted in Tables VI and XI is the improvement of the Organization recognition in relation to other corpora in Pt-BR: *Paramopama*'s NER classi-

TABLE III. SOME FEATURES USED BY THE NER CLASSIFIER

Feature	Description
Order=2	The number of things to the left that have to be cached to run the Viterbi algorithm: the maximum context of class features used.
Class	Puts a prior on the classes which is equivalent to how often the feature appeared in the training data.
PreviousWord	Gives a feature for previous word
Word	Gives a feature for actual word
NextWord	Gives a feature for next word
UseNGrams	Make features from letter n-grams, i.e., substrings of the word, with maximum N-Gram length equal 6
Subclassification	Convert the labeling of classes (but not the background) into a encodings (IO), I(nside), O(utside) 2-way classification for each class
UseGazettes	Use gazette features for (Person, Localization, Organization) tags
WordShape	Feature for shape forms of a word

TABLE IV. RESULTS FOR PERSON

Corpus	Precision Set 1	Recall Set 1	F-measure Set 1
First HAREM	76.32%	64.93%	70.16%
Second HAREM	82.55%	84.70%	83.61%
Full HAREM	70.03%	77.61%	73.63%
WikiNER	78.86%	58.77%	67.35%
Paramopama	86.72%	77.99%	82.12%
Paramopama + HAREM	80.44%	81.34%	80.89%

TABLE V. RESULTS FOR LOCATION (TEST SET 1)

Corpus	Precision Set 1	Recall Set 1	F1-Score Set 1
First HAREM	74.48%	71.24%	72.82%
Second HAREM	68.97%	89.97%	78.08%
Full HAREM	**81.16%**	74.92%	77.91%
WikiNER	49.34%	73.69%	59.10%
Paramopama	**71.43%**	88.63%	79.10%
Paramopama + HAREM	74.51%	88.96%	81.10%

TABLE VI. RESULTS FOR ORGANIZATION (TEST SET 1)

Corpus	Precision Set 1	Recall Set 1	F1-Score Set 1
First HAREM	68.61%	43.93%	53.56%
Second HAREM	76.99%	40.65%	53.21%
Full HAREM	**81.42%**	**42.99%**	**56.27%**
WikiNER	50.15%	40.66%	44.91%
Paramopama	**81.44%**	**63.55%**	**71.39%**
Paramopama + HAREM	89.93%	58.41%	70.82%

TABLE VII. RESULTS FOR TIME (TEST SET 1)

Corpus	Precision Set 1	Recall Set 1	F1-Score Set 1
First HAREM	**3.48%**	**2.55%**	**2.94%**
Second HAREM	88.74%	85.35%	87.01%
Full HAREM	65.94%	57.96%	61.69%
WikiNER	-	-	-
Paramopama	87.90%	87.90%	87.90%
Paramopama + HAREM	90.06%	92.36%	91.19%

TABLE VIII. OVERALL SCORE (TEST SET 1)

Corpus	Precision Set 1	Recall Set 1	F1-Score Set 1
First HAREM	63.32%	51.65%	56.89%
Second HAREM	77.18%	76.36%	76.77%
Full HAREM	74.64%	65.50%	69.77%
WikiNER	60.45%	59.44%	59.94%
Paramopama	**79.91%**	**79.66%**	**79.79%**
Paramopama + HAREM	**81.25%**	**80.30%**	**80.77%**

fier achieved an F-measure of almost 72% against 56% of HAREM's.

As presented in Tables VII and XII, First HAREM corpus had low performance classification for the Time entity in both test sets. The reason is that the words of this corpus tagged with Time entity only specify words representing dates. But in the test sets, there are many different temporal expressions that should also be tagged as Time. For example, in the sentence *"No primeiro dia do ano, Lula viajou para Brasília"* (*"On the first day of the year, Lula went to Brasilia."*), the expression *"No primeiro dia do ano"* (*"On the first day of the year"*) should be tagged as Time.

Some entities increase the chances for mislabeling. This is the case of Location and Organization. Depending on the context, a word may refer to a Location or an Organization. Precision values for Location class, for instance, show that

TABLE IX. RESULTS FOR PERSON (TEST SET 2)

Corpus	**Precision Set 2**	**Recall Set 2**	**F1-Score Set 2**
First HAREM	64.11%	54.90%	59.15%
Second HAREM	77.48%	82.11%	79.73%
Full HAREM	63.05%	67.01%	64.97%
WikiNER	83.16%	68.62%	75.19%
Paramopama	83.08%	76.30%	79.55%
Paramopama + HAREM	80.74%	78.42%	79.56%

TABLE X. RESULTS FOR LOCATION (TEST SET 2)

Corpus	**Precision Set 2**	**Recall Set 2**	**F1-Score Set 2**
First HAREM	75.68%	57.81%	65.55%
Second HAREM	77.40%	84.53%	80.81%
Full HAREM	**80.35%**	59.63%	68.46%
WikiNER	67.16%	91.30%	77.39%
Paramopama	**78.53%**	86.86%	82.49%
Paramopama + HAREM	81.69%	85.46%	83.54%

TABLE XI. RESULTS FOR ORGANIZATION (TEST SET 2)

Corpus	**Precision Set 2**	**Recall Set 2**	**F1-Score Set 2**
First HAREM	45.68%	37.67%	41.29%
Second HAREM	75.79%	46.40%	57.56%
Full HAREM	**50.87%**	**36.50%**	**42.50%**
WikiNER	76.05%	45.08%	56.61%
Paramopama	**70.26%**	**66.27%**	**68.21%**
Paramopama + HAREM	73.03%	62.81%	67.54%

Paramopama performed worse if compared to HAREM as highlighted in tables V and X.

IV. CONCLUSION

This paper presented *Paramopama*, a Named Entity corpus for Brazilian Portuguese (PtBR NE-tagged corpus). *Paramopama* is fully tagged with the following classes: Person, Location, Organization and Time. The paper has also presented a learned model for Paramopama and for some other known PtBR NE-tagged corpora, properly evaluated using Precision, Recall and F-Measure measures. All the resources such as

TABLE XII. RESULTS FOR TIME (TEST SET 2)

Corpus	**Precision Set 2**	**Recall Set 2**	**F1-Score Set 2**
First HAREM	5.00%	3.99%	4.44%
Second HAREM	89.02%	85.75%	87.36%
Full HAREM	65.52%	54.76%	62.65%
WikiNER	-	-	-
Paramopama	92.01%	90.86%	91.43%
Paramopama + HAREM	93.24%	93.41%	93.32%

TABLE XIII. OVERALL SCORE (TEST SET 2)

Corpus	**Precision Set 2**	**Recall Set 2**	**F1-Score Set 2**
First HAREM	55.07%	44.35%	49.13%
Second HAREM	79.40%	79.26%	79.33%
Full HAREM	68.12%	58.00%	62.65%
WikiNER	72.53%	75.90%	74.17%
Paramopama	**81.26%**	**81.90%**	**81.58%**
Paramopama + HAREM	**82.66%**	**82.02%**	**82.34%**

corpus, test sets and specifications are available[1].

The results show that an NER classifier learned from *Paramopama* performed slightly better than others built from state-of-the-art PtBR NE-tagged corpora, considering the evaluated classes. If we consider the Organization entity, Paramopama performed better in both test sets due mainly to the *Expansion* step of Paramopama.

We are continuously working on the expansion of *Paramopama* size in order to improve its coverage.

V. ACKNOWLEDGEMENT

The authors thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the financial support [Universal 14/2012, Processo 483437/2012-3] and for granting a scholarship to Carlos Mendonça and Fundação de Apoio à Pesquisa e Inovação Tecnológica do Estado de Sergipe (FAPITEC-SE) for granting a scholarship to Nayara Rosy.

REFERENCES

[1] The Stanford Natural Language Processing Group. (2015, April 10). Stanford Named Entity Recognizer(NER)[Online]. Available: http://nlp.stanford.edu/software/CRF-NER.shtml.

[2] D. O. F. Do Amaral, and R. Vieira, NERP-CRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de Conditional Random Fields. Linguamática, vol. 6, no. 1, pp. 41-49, 2014.

[1]https://goo.gl/9e3O1O

[3] N. Chinchor, MUC-7 Named Entity Task Definition (version 3.5), in Proceedings of the 7th Message Understanding Conference, 1998, Appendix E.

[4] D. O. F. Do Amaral, "O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa", Ph.D. dissertation, Pontifícia Universidade Católica do Rio Grande do Sul, 2013.

[5] C. Freitas, C. Mota, D. Santos, H. G. Oliveira, and P. Carvalho, Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese, in LREC, 2010.

[6] D. Jurafsky, and J. H. Martin, Speech & language processing, Pearson Education India, 2000.

[7] J. Lafferty, A. Mccallum, and F. C.N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, 2001.

[8] ACE (Automatic Content Extraction) Chinese Annotation Guidelines for Entities, Linguistic Data Consortium, Version 5.6.6, https://goo.gl/rgcCpm, 2006.

[9] A. McCallum, and W. Li, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, in Proceedings of the seventh conference on Natural language learning at HLT-NAACL, 2003, Volume 4. Association for Computational Linguistics, 2003, pp. 188-191.

[10] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, Learning multilingual named entity recognition from Wikipedia. Artificial Intelligence, v. 194, p. 151-175, 2013.

[11] D. Santos, and N. Cardoso, A golden resource for named entity recognition in Portuguese, in Computational Processing of the Portuguese Language, Springer Berlin Heidelberg, 2006, pp. 69-79.

[12] B. Settles, Biomedical named entity recognition using conditional random fields and rich feature sets, in Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Association for Computational Linguistics, 2004, pp. 104-107.

[13] E. F. T. K. Sang, and F. Meulder, Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition, in Proceedings of the 6th Conference on Natural Language Learning, Taipei-Taiwan, 2002, pp. 1–4.

[14] E. F. T. K. Sang, and F. Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, Association for Computational Linguistics, 2003, pp. 142-147.

[15] R. Weischedel, and A. Brunstein, BBN pronoun coreference and entity type corpus, Linguistic Data Consortium, Philadelphia, Tech Rep. LDC2005T33, 2005.