STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

COREFERENCE RESOLUTION FOR PORTUGUESE USING PARALLEL CORPORA WORD ALIGNMENT

JOSÉ GUILHERME CAMARGO DE SOUZA AND CONSTANTIN ORĂSAN

1. Introduction

The field of Information Extraction (IE) studies and creates techniques for turning the unstructured information present in natural language texts into structured data [7]. An important part of this process is coreference resolution, the task which identifies when different noun phrases refer to the same discourse entity in a text. Coreference resolution is a field which has been extensively researched for English (see [9] for a comprehensive overview of methods), but received less attention for other languages. This is due to the fact that the vast majority of the existing methods are based on machine learning and therefore require extensive annotated data.

The aim of this paper is to present a system that automatically extracts coreference chains from texts in Portuguese without having to resort to Portuguese corpora manually annotated with information about coreferential links. To achieve this goal, it is necessary to implement a method which can automatically obtain data that can be used for training a supervised machine learning coreference resolver for Portuguese. In this work, the training data is acquired by using an English-Portuguese parallel corpus in which the coreference chains annotated in the English part of the corpus are projected to the Portuguese part of the corpus. This approach is similar to the one proposed by [13] for projecting coreference chains from English to Romanian. In contrast to the method developed by [13], our goal here is not to create an annotated resource, but to implement a fully functional coreference resolver for Portuguese.

The rest of this paper presents the current stage of the development of the system and is structured as follows: Section 2 presents a brief overview of relevant work in coreference resolution with emphasis on Portuguese. Section

Received by the editors: April 25, 2011.

2010 *Mathematics Subject Classification.* 68T50.

1998 *CR Categories and Descriptors.* I.2.7 [**Natural Language Processing**]: Discourse – *Coreference Resolution*.

Key words and phrases. coreference resolution, parallel corpus, machine learning.

3 presents the approach proposed in this paper. Preliminary evaluation results are presented in Section 4 and briefly discussed.

2. Related Work

Despite attempts to develop rule-based coreference resolution systems as part of the MUC competitions or using the MUC data [5], most of the existing systems rely on machine learning approaches [9]. This is possible for English where there are several annotated corpora large enough to be used for training, but not for languages such as Portuguese which lacks the necessary resources. As a result, most work for Portuguese focused on certain types of pronominal anaphora resolution [12, 4] or problems related to coreference and anaphora resolution such as anaphoricity classification [3]. The only available corpus annotated with coreferential data is the Summ-It corpus [2] and to the best of our knowledge the only work that uses it for the development of a machine learning method for coreference resolution is [17]. Due to the small size of the corpus, the evaluation results presented in that paper are below the ones obtained by state-of-the-art systems for English.

The Summ-It corpus imposes restrictions on which supervised machine learning approaches that can be used. For example, it is not possible to use a large list of features, each with a detailed set of attributes as suggested by some researchers, because this requires a large quantity of data for training. Summ-It is not a big corpus, it contains around 700 coreferential expressions distributed in 50 newswire texts. This is not as large as the corpora normally used to train machine learning approaches for other languages such as English (MUC[1] and ACE[2]) and Spanish (AnCora [15]).

The next section presents a method that does not rely on the availability of manually annotated data for coreference in the language in which the text is processed.

3. The method

As previously mentioned, the goal of this research is to extract coreference chains automatically from Portuguese texts. To achieve this, a parallel corpus is employed to project coreference relations from the English part of the corpus to the Portuguese part. The relations projected are then used for training a supervised machine learning model capable of identifying coreference chains in Portuguese. Figure 1 shows an overview of the whole system and each step depicted in the figure is described next.

[1] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/muc_data/muc_data_index.html

[2] http://projects.ldc.upenn.edu/ace/data/

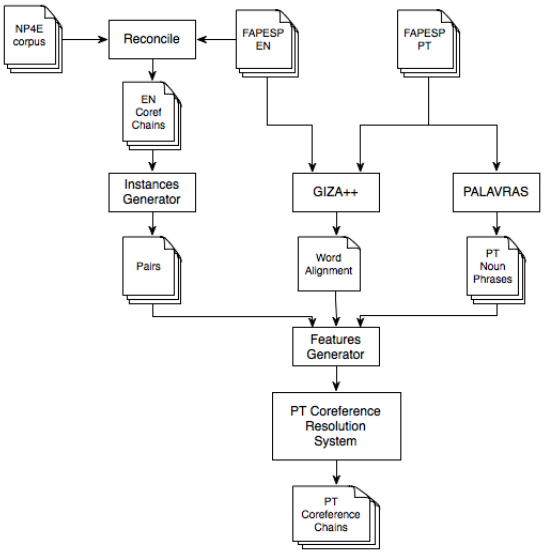

FIGURE 1. The overall structure of the system

3.1. **Annotation of English coreference chains.** The first step of the processing identifies coreference chains in the English part of the parallel corpus. An off-the-shelf machine learning based coreference resolver for the English language, Reconcile [18], is used to automatically annotate the text with coreference chains. The authors of Reconcile report the results in terms of MUC score and B^3 score. The MUC F-Measure reported is of 68.50 for the MUC6 dataset and 62.80 for the MUC7 dataset. The B^3 F-Measure is of 70.88 for the MUC6 dataset and of 65.86 for the MUC7 dataset.

3.2. **Generation of English pairs.** The automatic annotation obtained in the previous step is used to generate pairs of expressions (antecedent and anaphor) that can be projected in the Portuguese corpus and used to train a machine learning algorithm. The algorithm implemented for generating the positive pairs (anaphoric pairs) always chooses the most confident antecedent for a given anaphor as proposed in [10]. For each non-pronominal noun phrase, it is assumed that the most confident antecedent is the closest non-pronominal preceding antecedent. For generating the negative pairs (non-anaphoric pairs), the algorithm implemented is the one proposed by [16]. The negative pairs are formed by using expressions that occur in between the expressions in the

positive chains. The anaphor is always an expression that belongs to a coreference chain and the antecedent is an expression that does not belong to the same chain or that does not belong to any chain.

3.3. **Identification of the NP.** This step identifies NPs in both languages because they correspond to mentions and therefore can be in coreference relations. It needs to be performed explicitly only for Portuguese, as NPs in the English part of the corpus are identified during the coreference resolution process. The Portuguese NPs are identified using the Constraint Grammar based parser PALAVRAS [1]. The authors report 99% accuracy for part-of-speech tagging and about 97% accuracy for syntactic function detection.

3.4. **Word-by-word alignment.** Even though the method proposed in this paper relies on a parallel corpus, most parallel corpora do not have a word-by-word alignment as is required in the next step. For this reason, Giza++ [11] is used to produce this alignment.

3.5. **Generation of Portuguese training examples.** The word-by-word alignment is used in the process of generating Portuguese training examples from English pairs. Given the errors introduced by the identification of English NPs and by the alignment process, the English NPs are not directly mapped to Portuguese NPs. Instead, a matching algorithm is used to identify the best Portuguese NP to be aligned with the English NP. Once a pair is identified in the Portuguese data, features are extracted in order to produce training examples. The features used are a mix of the ones proposed by [16] and [14]. These features are used to train a machine learning model that identifies coreferential chains in Portuguese.

4. Evaluation

4.1. **Corpus.** Our method was evaluated on an English-Portuguese parallel corpus which contains texts from the *Revista Pesquisa FAPESP*[3] (FAPESP Research Magazine). The corpus contains 646 texts with a total of 17427 sentences. The English part contains around 464.000 words, and there are about 433.000 words in the Portuguese part.

In addition, the NP4E corpus [6] is used internally by Reconcile to learn the coreference model and the success of the proposed methodology is assessed on the Summ-it corpus [2]. All three corpora contain newswire texts which makes them comparable to a certain extent. However, due to the differences between them, the results may not be as high as they would be if only one corpus

[3] `http://revistapesquisa.fapesp.br/`

was used (e.g. if the parallel corpus was used both for training the English coreference model and to evaluate the Portuguese coreference resolver).

4.2. **Evaluation results.** For the FAPESP corpus, the system generated 94,990 coreference chains. Most of these chains are singleton (i.e. chains formed by only one expression): 82,272. This represents approximately 86% of the expressions. The remaining 14% are chains formed by two or more expressions. Using these chains, our current system generates 21,849 positive pairs (approximately 4.7%) and 436,033 negative pairs (approximately 95.2%) out of 457,882 pairs.

These pairs were projected from one side of the corpus to the other using the current implementation of the projection algorithm. The algorithm projected 3,569 positive pairs (approximately 7.6%) and 43,174 (approximately 92.3%) out of 46,543 projected pairs.

The coreference chains extracted by the system were scored using two scoring metrics, MUC [19] and CEAF [8]. The F-Measure values are 7.12 for the MUC score and 14.37 for the CEAF score. The baseline implemented clusters all the pairs of mentions that share the same head word. The performance is identical to the baseline. Analysis of the results reveals that most of the projected coreferential pairs used for training also have head matching. This is due to the fact that the chains extracted by Reconcile contain lots of pairs which have the same head. This phenomenon is intensified by the projection algorithm.

References

[1] E Bick. *The parsing system PALAVRAS: automatic grammatical analysis of Portuguese in a constraint grammar framework*. Phd, Arhus, 2000.

[2] Sandra Collovini, Thiago I Carbonel, Juliana Thiesen Fuchs, and Renata Vieira. Summit: Um corpus anotado com informacoes discursivas visando à sumarizacao automática. In *TIL - V Workshop em Tecnologia da Informação e da Linguagem Humana*, pages 1605–1614, Rio de Janeiro, 2007.

[3] Sandra Collovini and Renata Vieira. Learning Discourse-new References in Portuguese Texts. In *TIL 2006*, pages 267–276, 2006.

[4] R.R.M. Cuevas and Ivandré Paraboni. A Machine Learning Approach to Portuguese Pronoun Resolution. *Proceedings of the 11th Ibero-American conference on AI: Advances in Artificial Intelligence*, pages 262–271, 2008.

[5] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).*, pages 168 – 175, July 2002.

[6] Laura Hasler, Constantin Orăsan, and Karin Naumann. NPs for Events: Experiments in Coreference Annotation. pages 1167 – 1172, Genoa, Italy, May, 24 – 26 2006.

[7] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall, 2nd edition, 2009.

[8] Xiaoqiang Luo. On coreference resolution performance metrics. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, 2005.

[9] Vincent Ng. Supervised Noun Phrase Coreference Research : The First Fifteen Years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, number July, pages 1396–1411, 2010.

[10] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, Philadelphia, United States, 2002.

[11] Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.

[12] Ivandré Paraboni and Vera Lúcia Strube De Lima. Possessive Pronominal Anaphor Resolution in Portuguese Written Texts - Project Notes. In *17th International Conference on Computational Linguistics (COLING-98)*, pages 1010–1014, Montreal, Quebec, Canada, 1998. Morgan Kaufmann Publishers.

[13] Oana Postolache, Dan Cristea, and Constantin Orasan. Transferring Coreference Chains through Word Alignment. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.

[14] Marta Recasens and Eduard Hovy. A deeper look into features for coreference resolution. *Anaphora Processing and Applications*, (i):29–42, 2009.

[15] Marta Recasens and M. Antònia Martí. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):341–345, 2009.

[16] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544, December 2001.

[17] José Guilherme Camargo De Souza, Patricia Nunes Gonçalves, and Renata Vieira. Learning Coreference Resolution for Portuguese Texts. In António Teixeira, Vera Lúcia Strube De Lima, Luís Caldas De Oliveira, and Paulo Quaresma, editors, *Computational Processing of the Portuguese Language - 8th International Conference, PROPOR 2008*, pages 153–163, Aveiro, Portugal, 2008. Springer Berlin / Heidelberg.

[18] Veselin Stoyanov, Claire Cardie, Nathan Gilbert, and David Buttler. Coreference Resolution with Reconcile. In *Proceedings of the Joint Conference of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*. Association for Computational Linguistics, 2010.

[19] Marc Vilain, John Burger, John Aberdeen, and Dennis Connolly. A model-theoretic coreference scoring scheme. *Proceedings of the 6th conference on Message understanding*, pages 45–52, 1995.

Research Institute in Information and Language Processing, University of Wolverhampton, United Kingdom

E-mail address: joseguilhermecs@gmail.com and C.Orasan@wlv.ac.uk