Can Projected Chains in Parallel Corpora Help Coreference Resolution?

José Guilherme Camargo de Souza and Constantin Orăsan

Research Group in Computational Linguistics,
University of Wolverhampton,
Wolverhampton, UK
jose.camargo.souza@gmail.com, C.Orasan@wlv.ac.uk

Abstract. The majority of current coreference resolution systems rely on annotated corpora to train classifiers for this task. However, this is possible only for languages for which annotated corpora are available. This paper presents a system that automatically extracts coreference chains from texts in Portuguese without the need for Portuguese corpora manually annotated with coreferential information. To achieve this, an English coreference resolver is run on the English part of an English-Portuguese parallel corpus. The coreference pairs identified by the resolver are projected to the Portuguese part of the corpus using automatic word alignment. These projected pairs are then used to train the coreference resolver for Portuguese. Evaluation of the system reveals that it does not outperform a head match baseline. This is due to the fact that most of the projected pairs have the same head, which is learnt by the Portuguese classifier. This suggests that a more accurate English coreference resolver is necessary. A better projection algorithm is also likely to improve the performance of the system.

Keywords: coreference resolution, parallel corpus, machine learning.

1 Introduction

Coreference resolution is "the problem of identifying which noun phrases (NPs) or mentions refer to the same real-world entity in a text or dialogue" [16]. This is a very important task for many NLP fields including information extraction, question answering, automatic summarisation and machine translation. The task has been extensively researched for English (see [17] for an overview), but has received less attention for other languages. This is largely due to the fact that most methods require corpora annotated with coreferential information which are not available for many languages.

This paper presents a system that automatically extracts coreference chains from texts in Portuguese without the need for Portuguese corpora manually annotated with coreferential information. In order to achieve this, a method is developed for automatically producing training data for a supervised machine learning coreference resolver for Portuguese. This is done by using an English-Portuguese parallel corpus in which the coreference chains annotated in the

I. Hendrickx et al. (Eds.): DAARC 2011, LNAI 7099, pp. 59–69, 2011.

English part are projected to the Portuguese part in a similar way to that proposed by [21] for projecting coreference chains from English to Romanian. In contrast to the method developed by [21], the goal of our method is not to create an annotated resource, but to implement a fully functional coreference resolver for Portuguese. It should be pointed out that there is nothing in the overall idea that makes it specific to the English-Portuguese language pair. The method can be applied to other language pairs as long as there is a parallel corpus available and the components that deal with Portuguese are replaced with the corresponding components for the target language.

The remainder of this paper presents and evaluates the system and is structured as follows: A brief overview of related research is presented in Section 2, followed by Section 3 which describes the approach proposed in this work. Evaluation results are presented and discussed in Section 4. The paper finishes with conclusions in Section 5.

2 Related Work

As in many other NLP fields, two main approaches are used in coreference resolution: knowledge engineering methods and machine learning methods. The knowledge engineering methods generally require humans to manually create rules which determine whether two noun phrases are coreferential or not. These methods usually exploit regularities of the documents they process and are designed for specific applications [12]. Given the difficulty of creating rules manually, the vast majority of existing systems use machine learning approaches for this task [17]. The most common model used for this is the *mention-pair* model where the system is first trained to classify whether pairs of noun phrases are coreferential or not [1,14]. In the second step, a clustering algorithm is used to group entities into coreferential chains. For the classification stage, but sometimes also for the clustering stage, positive and negative instances are extracted from an annotated corpus and used to train the machine learning algorithm. Even for the other models used in coreference resolution, an annotated corpus is necessary. In light of this, it becomes obvious that the lack of availability of annotated data is the main bottleneck in the development of machine-learning based coreference resolution systems.

Some languages have enough annotated data to allow training of machine learning methods. For example for English, the MUC[1] and ACE[2] corpora have been successfully used by many researchers. In recent years, corpora annotated with coreference containing Spanish and Catalan [23], and Dutch [11] texts were also released, in this way facilitating the development of coreference resolution systems for these languages. Unfortunately, for many other languages such corpora are either not available or are rather small to allow training of robust methods.

[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/`

One such language is Portuguese, for which, to the best of our knowledge, the only available corpus annotated with coreferential data is the Summ-It corpus [7] and the only study that uses it to develop a supervised machine learning approach for coreference resolution is [25]. The small size of the corpus imposes some limits on the kind of machine learning that can be used. So far, most of the work for Portuguese has focused on certain types of pronominal anaphora ([20], [5] and [9]) or problems related to coreference and anaphora resolution such as anaphoricity classification [8]. This was the main motivation for developing the proposed method.

As mentioned before, our research was inspired by [21] but it goes several steps further. The goal of the work presented there is to provide a bootstrapping method for creating coreferentially annotated data. For this reason, several manual operations are envisaged in the process. The starting point is an English-Romanian parallel corpus in which the English texts were manually annotated with coreference information. In contrast, we assume no annotation available for English and use an English coreference resolver to obtain coreference pairs. The annotation is transferred to Romanian using an automatic aligner, but it is not used to train a system. Instead it is foreseen that it is given to human annotators for post-editing. In this paper, we intend to use the transferred information to train a coreference resolver for Portuguese.

Several researchers have used parallel corpora and projection of information across languages before: [29] project part-of-speech tags onto English-French and English-Chinese corpora; [15] improve the performance of anaphora resolvers for English and French using an English-French parallel corpus; [2] project word senses onto an English-Italian parallel corpus; and [19] use projection to reduce the effort on annotation of semantic roles and presents evaluation on an English-German parallel corpus for both automatically and manually annotated English data.

3 Methodology

The aim of our research is to extract coreference chains automatically from Portuguese texts without the need for an annotated corpus for Portuguese. To achieve this, the system presented in Figure 1 is used.

The system is composed of several components that can be roughly grouped into three main modules: automatic corpus annotation (English coreference resolution and Portuguese parsing and noun phrase extraction), alignment (word alignment of the parallel corpus), and coreference resolution for Portuguese (instance and feature generation, and coreference resolution). This section gives a brief overview of the system. More details about the architecture used can be found in [10]. The section finishes with some observations about adapting the system to other language pairs.

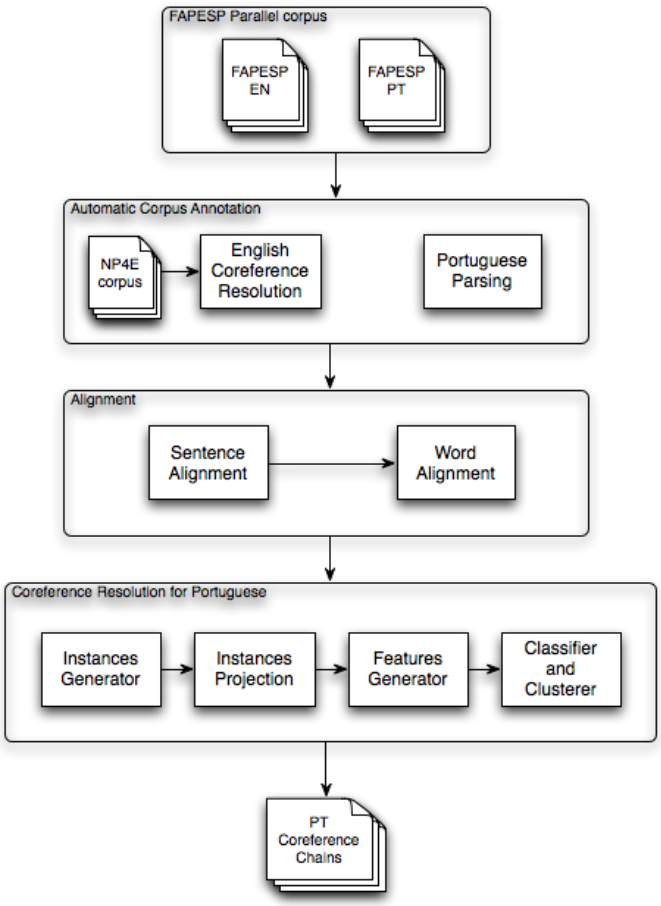

Fig. 1. The overview of the system's architecture

3.1 Automatic Corpus Annotation

The first module automatically annotates the corpus with linguistic information required at later stages of processing. Reconcile [26] is used to annotate the coreferential links in the English side of the corpus. Reconcile is an off-the-shelf coreference resolution system for English that uses machine learning and was trained on a collection of texts from the Reuters corpus. The system was run as it is and no attempts were made to tune it for our texts. As a result of using the coreference resolver, the noun phrases in the English texts are also annotated.

The Portuguese side of the corpus is annotated with morpho-syntactic information using the PALAVRAS parser [3]. This information includes boundaries of NPs and features such as the head of the NP, number and gender information, and the type of NP, which are used when the Portuguese coreference chains are identified.

3.2 Alignment

The noun phrases identified by the previous module are used by the GIZA++ word aligner [18] to establish mappings between the noun phrases in the two sides of the corpus. This step is necessary because most parallel corpora do not have a word-by-word alignment. Because the word aligner requires that the corpus is aligned at sentence level, we run the TCAlign sentence aligner [4] first.

3.3 Coreference Resolution for Portuguese

The last module of the system is responsible for performing the actual coreference resolution. This process consists of the following three steps: instance generation and projection, feature generation and classification and clustering.

Instance Generation and Projection: The idea of the method is to project the coreferential links from English to Portuguese and use them to train a coreference resolver for Portuguese. Given the errors introduced by the NP extractors and by the word-to-word aligner, it is not possible to directly map English NPs to Portuguese NPs. Instead, for each English anaphor and antecedent pair, the algorithm checks whether their heads have a corresponding word in Portuguese. If so, the pair is projected to Portuguese and used as a positive pair to train the Portuguese coreference resolver. For generating negative pairs (non-anaphoric pairs), the anaphoric expression projected from English is paired with all the Portuguese noun phrases which occur between the projected positive pair. This is the standard in coreference resolution [24], but it generates training data which contain a much larger proportion of negative instances.

Feature Generation: After the pairs are produced for Portuguese, a set of features is extracted for each of the pairs to be used by the learning algorithm. The features were inspired by previous work on English and Portuguese coreference resolution [24,25,22] and contain a mixture of surface-based features (e.g. head match and substring match), semantic information (e.g. number and gender agreement, and type of entity associated by PALAVRAS) and syntactic information (e.g. whether the antecedent or anaphor are the subject of the sentence). A more detailed description of the 11 features used in our research can be found in [10].

Classification and Clustering: The features extracted in the previous step are used to produce training instances for a machine learning algorithm. After experimenting with several algorithms, we decided to use JRip, WEKA's implementation of decision rules. Using the induced rules, each instance is classified as coreferential or non-coreferential. The classified pairs are clustered together using the Closest-First clustering algorithm, in this way producing the coreferential chains.

3.4 Adaptation to Other Languages

The system described in this section was developed for the English-Portuguese language pair, but it can be easily adapted for other languages as long as they have a parallel corpus and the necessary annotation tools. Given the availability of English coreference resolvers, the assumption is that the coreference links will be projected from English, but even this can be changed if a coreference resolver for the source language is available.

In order to develop a coreference resolver for a language other than Portuguese, it is necessary to have an NP extractor for that language, which also provides information about the heads of NPs. If necessary this can be replaced by an NP chunker and a set of heuristics which can approximate the head of the NP. The word alignment algorithm used here can be applied to any language pairs as long as there is a language model for that language pair. The pair projection method proposed here can also be used without modifications, but depending on the language pair and the accuracy of the NP identifiers used, improvements can be brought to the algorithm. The classifier used for Portuguese to decide whether a pair is coreferential or not can be used for other languages as well, but the set of features may need to be changed depending on the characteristics of the language.

4 Evaluation

The system described in the previous section does not depend on a particular parallel corpus. In this section, we explain first how we instantiated the system using the FAPESP parallel corpus. The performance of this instance of the system was evaluated using the Summ-It corpus. The section finishes with an error analysis.

4.1 Instantiation of the System

As explained earlier, the parallel corpus is central to the method developed here. In order to check the performance of the developed system, we used the English-Portuguese parallel corpus extracted from the electronic version of the *Revista Pesquisa FAPESP* Brazilian magazine[3]. The corpus contains 646 texts about domestic and international scientific policy, and about research carried out in Brazil and other countries. The English side contains around 464,000 words, and there are approximately 433,000 words in the Portuguese side.

For the FAPESP corpus, Reconcile generated 94,990 coreference chains in the English part. 82,272 of these chains are singleton (i.e. chains formed by only one expression) which is approximately 86% of the expressions identified in the text. The remaining 14% are chains formed by two or more expressions.

Using the chains extracted by Reconcile, the system generated 21,849 positive pairs (approximately 4.8%) and 436,033 negative pairs (approximately 95.2%)

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

out of 457,882 pairs. The English pairs are projected using the projection algorithm described in section 3. The algorithm successfully projected 3,569 positive pairs (7.6%) and 43,174 negative pairs (92.4%).

The increase in the proportion of positive pairs is explained by the way the pairs are projected. The projection algorithm uses the heads of noun phrases and the sentences where the noun phrases occur. If any of this information is not available (due to problems in syntactic annotation or in the alignment) or if the process of finding the expressions in the aligned corpus fails, the algorithm ignores the instance.

These projected pairs are used to train a supervised machine learning model that is used further on as a classifier in the clustering algorithm. The algorithm used to train the model is the WEKA toolkit [28] implementation of the decision rules [6] algorithm (JRip).

4.2 Evaluation of the System

The system was evaluated on the Summ-It corpus [7], a coreferentially annotated corpus that contains around 17,000 words and 700 coreferential expressions distributed in 50 newswire texts. This corpus was used because it was not employed in the development of the system. The performance of the coreference resolver is scored using the MUC [27] and CEAF [13] measures.

In order to have a better understanding of the performance of the system, a baseline was developed as well. The baseline classifies two expressions as coreferential if they have the same head. The coreferential chains are built using the same Closest-First clustering algorithm used by our system.

The baseline and our system obtained the same scores on the Summ-it corpus: 7.12% MUC f-measure and 14.37% CEAF f-measure. One of the reasons why the MUC score has a lower f-measure is because it penalises missed links and several chains extracted present only part of the expressions they should contain. The next section contains an error analysis which tries to explain the results.

4.3 Error Analysis

Our system relies on several components which address difficult language processing problems and which, even though they represent the state-of-the-art in the field, are still not highly accurate. As a result, each of these components introduces errors which propagate throughout the system, contributing to its low performance. This section focuses on the coreference resolution components in an attempt to understand the low accuracy of the proposed system.

Error Analysis of Output of Reconcile: The authors of Reconcile reported MUC f-measure scores of 68.50% for the MUC-6 corpus and 62.80% for the MUC-7 corpus. In order to assess the accuracy of Reconcile on our data, five texts from the FAPESP corpus with approximately 3,600 words and 846 markables in the English side were annotated with coreferential information. Taking the manual

(sys) ⟨ *the old light microscope*, *the electron microscope*, *the electron microscope in Germany*, *the electron microscope* ⟩

(ref_a) ⟨ *the electron microscope*, *The electron microscope*, *The electron microscope*, *an electron microscope*, *the electron microscope*, *an electron microscope*, *this instrument* ⟩

(ref_b) ⟨ *the light microscope*, *The light microscope* ⟩

Fig. 2. Chain extracted by Reconcile and its corresponding manual annotation

annotation as reference, Reconcile's output obtained a MUC f-measure of 76.8%. Given the small size of the corpus, we cannot argue that Reconcile's performance on our texts is significantly better than the one obtained on the MUC dataset. However, we believe that it indicates that its performance on our texts is at least in line with that reported by the authors of Reconcile.

Comparing the system annotation and the manual annotation, it is possible to confirm the intuition that many entities in the chains, albeit sharing the same head, belong to different chains. One example of such an error is presented in figure 2. The chain sys is the output produced by Reconcile. The chains ref_a and ref_b are two different chains identified by our annotator. The former is about an electron microscope and the latter is about a light microscope. In the sys chain Reconcile mixed elements from the two chains ref_a and ref_b into one long chain.

The same phenomenon was observed in other chains extracted by the system in this set of 5 texts. It is possible to conclude that longer chains contain more expressions that do not belong to them, generating undesired noise that is projected to the Portuguese side of the corpus.

Given Reconcile's bias towards using head match for classifying entities as coreferential, we analysed all the non-singleton chains identified in the FAPESP corpus. Comparison of all the expressions in the chain in a pairwise fashion revealed that about 53% of the pairs share the same head. Among this 53%, there are a fair number of errors where entities share the same head, but do not refer to the same entity as in the examples above.

Evaluation of the Learnt Rules: The JRip algorithm was run with 10-fold cross-validation and default parameters. The automatically induced classifier correctly classified 45,944 out of 46,743 instance pairs projected (approximately 98%). However, most of the instances fall under only one rule that uses a feature that verifies if the heads of both expressions match. Analysis of the 3,569 coreferent pairs showed that 2,978 (approximately 83%) of them have the same head. This leaves only 591 pairs (17%) that are positive but that do not have the same head. The small amount of positive and non-head matching pairs is not informative enough to help the JRip algorithm employ features other than the head match feature.

Figure 3 shows the only 2 rules induced by the system. The first rule classifies two noun phrases as coreferential if they have the same head and it is applied in the majority of cases. The second rule is applied only in five instances, which explains why our method and the baseline obtain the same results.

```
if (head_match = 1) => class=C
if (number_agrmt = 1) and (ant_appos = 1) and
   (sem_class_agrmt = 1) and (word_overlap = point5) and
   (ana_appos = 0) => class=C
else => class=NC
```

Fig. 3. The rules generated by the JRip algorithm

5 Conclusions and Future Work

This paper presented a system which extracts coreference chains from Portuguese texts without having to resort to Portuguese corpora manually annotated with coreferential information. The system implements a method that automatically obtains data for training a supervised machine learning coreference resolver for Portuguese.

The training data is generated by using an English-Portuguese parallel corpus from which the coreference chains annotated in the English part of the corpus are projected to the Portuguese part of the corpus. The coreference chain extraction system for Portuguese was tested in a corpus annotated with coreference chains in Portuguese. The results of the system on this corpus are comparable to the baseline.

The results of the system described here are strongly influenced by the coreference links identified in the English part of the corpus as the errors generated there are propagated throughout the pipeline. Therefore, the use of a better performing English coreference resolution system might improve the overall performance of the Portuguese resolver. A way to filter out errors introduced by the English coreference resolver is to use gender information from the Portuguese part to identify and remove these errors in a similar manner to what [15] did for pronoun resolution.

The projection algorithm used to transfer pairs has a strong influence on the accuracy of the Portuguese coreference resolver. Different methods for performing the projection might be implemented and tested. As future work, an evaluation of the projected pairs should be carried out in order to evaluate the strong points and the pitfalls of the algorithm employed.

An alternative route for developing the system is not to use a parallel corpus, but instead to automatically translate documents to the target language. This approach could prove useful for languages where there is not enough parallel data, but is likely to introduce additional problems due to the errors introduced by the machine translation step.

Acknowledgements. iThis work was partially supported by the European Commission, Education & Training, Erasmus Mundus: EMMC 2008-0083, Erasmus Mundus Masters in NLP & HLT programme. The authors would like to thank Sheila Castilho Monteiro de Sousa for her help in the annotation process.

References

1. Aone, C., Bennett, S.W.: Evaluating automated and manual acquisition of anaphora resolution strategies. In: The 33rd Annual Meeting on Association for Computational Linguistics, pp. 122–129 (1995)
2. Bentivogli, L., Pianta, E.: Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. Natural Language Engineering 11(03), 247 (2005), http://www.journals.cambridge.org/abstract_S1351324905003839
3. Bick, E.: The parsing system PALAVRAS: automatic grammatical analysis of Portuguese in a constraint grammar framework. Phd, Arhus (2000)
4. Caseli, H.D.M.: Alinhamento sentencial de textos paralelos português-inglês. Master thesis, USP (2002), http://www2.dc.ufscar.br/~helenacaseli/pdf/2002/QualiMestrado.pdf
5. Chaves, A., Rino, L.: The Mitkov Algorithm for Anaphora Resolution in Portuguese. In: The 8th International Conference on Computational Processing of the Portuguese Language, p. 60 (2008)
6. Cohen, W.: Fast effective rule induction. In: 12th International Workshop Conference on Machine Learning, pp. 115–123. Morgan Kaufmann Publishers, Inc. (1995)
7. Collovini, S., Carbonel, T.I., Fuchs, J.T., Vieira, R.: Summ-it: Um corpus anotado com informacoes discursivas visando à sumarizacao automática. In: TIL - V Workshop em Tecnologia da Informação e da Linguagem Humana, Rio de Janeiro, pp. 1605–1614 (2007)
8. Collovini, S., Vieira, R.: Learning Discourse-new References in Portuguese Texts. In: TIL 2006, pp. 267–276 (2006)
9. Cuevas, R., Paraboni, I.: A Machine Learning Approach to Portuguese Pronoun Resolution. In: The 11th Ibero-American Conference on AI: Advances in Artificial Intelligence, pp. 262–271 (2008)
10. de Souza, J., Orăsan, C.: Coreference resolution for Portuguese using parallel corpora word alignment. In: The International Conference on Knowledge Engineering, Principles and Techniques (KEPT 2011), Cluj-Napoca, Romania (July 2011)
11. Hoste, V., Pauw, G.D.: KNACK-2002: a Richly Annotated Corpus of Dutch Written Text. In: The Fifth International Conference on Language Resources and Evaluation, pp. 1432–1437. ELRA (2006)
12. Konstantinova, N., Orăsan, C.: Issues in topic tracking in wikipedia articles. In: The International Conference on Knowledge Engineering, Principles and Techniques (KEPT 2011), Cluj-Napoca, Romania, July 4-6 (2011)
13. Luo, X.: On coreference resolution performance metrics. In: The Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 25–32 (2005)
14. McCarthy, J.F., Lehnert, W.G.: Using Decision Trees for Coreference Resolution. In: International Joint Conference on Artificial Intelligence, pp. 1050–1055 (1995)

15. Mitkov, R., Barbu, C.: Using bilingual corpora to improve pronoun resolution. Languages in contrast 4(2), 201–212 (2004)
16. Ng, V.: Graph-Cut-Based Anaphoricity Determination for Coreference Resolution. In: NAACL 2009: Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 575–583. Association for Computational Linguistics, Boulder (2009)
17. Ng, V.: Supervised Noun Phrase Coreference Research: The First Fifteen Years. In: ACL 2010, pp. 1396–1411 (July 2010)
18. Och, F.J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29(1), 19–51 (2003)
19. Padó, S., Lapata, M.: Cross-lingual annotation projection of semantic roles. J. Artificial Intelligence Research. 36, 307–340 (2009)
20. Paraboni, I., Lima, V.L.S.D.: Possessive Pronominal Anaphor Resolution in Portuguese Written Texts - Project Notes. In: 17th International Conference on Computational Linguistics (COLING 1998), pp. 1010–1014. Morgan Kaufmann Publishers, Montreal (1998)
21. Postolache, O., Cristea, D., Orăsan, C.: Transferring Coreference Chains through Word Alignment. In: The 5th International Conference on Language Resources and Evaluation, Genoa, Italy (2006)
22. Recasens, M., Hovy, E.: A deeper look into features for coreference resolution. Anaphora Processing and Applications (i), 29–42 (2009)
23. Recasens, M., Martí, M.A.: AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. Language Resources and Evaluation 44(4), 341–345 (2009)
24. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics 27(4), 521–544 (2001)
25. 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)
26. Stoyanov, V., Cardie, C., Gilbert, N., Buttler, D.: Coreference Resolution with Reconcile. In: The Joint Conference of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010). Association for Computational Linguistics (2010)
27. Vilain, M., Burger, J., Aberdeen, J., Connolly, D.: A model-theoretic coreference scoring scheme. In: The 6th Conference on Message Understanding, pp. 45–52 (1995)
28. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann (2005)
29. Yarowsky, D., Ngai, G.: Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In: Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies 2001, pp. 1–8. Association for Computational Linguistics, Pittsburgh (2001)