Machine Learning Approaches to Coreference Resolution

Nguy G. L.

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czech Republic.

Abstract. This paper introduces three machine learning approaches to noun phrase coreference resolution. The first of them gives a view of coreference resolution as a clustering task. The second one applies a noun phrase coreference system based on decision tree induction and the last one experiments with using the Bell tree to represent the search space of the coreference resolution problem. The knowledge gained from these experiments can be conducive to development of a Czech coreference resolution system.

1. Introduction

In the spoken and written language it is commonly observed that the same real-world entity is refered to by a variety of noun phrases (NPs). The task of noun phrase coreference resolution is to determine which noun phrases in a text or dialogue refer to the same real-world entity. An accurate noun phrase coreference resolution is required by many natural language processing applications such as machine translation, information extraction etc.

This paper presents three algorithms for noun phrase coreference resolution. The first algorithm is based on the view of the problem as one of clustering the noun phrases (Section 3). Using a description of each noun phrase and a method for measuring the distance between two noun phrases, the clustering algorithm partitions the noun phrases into equivalence classes [*Cardie and Wagstaff*, 1999].

Decision tree algorithm was applied for coreference resolution by *Ng and Cardie* [2002] (Section 4). They used the C4.5 decision tree induction system [*Quinlan*, 1993] to train a classifier that decides whether or not two noun phrases in a document are coreferent. A clustering algorithm then constructs a partition on all noun phrases.

Section 5 outlines how *Luo et al.* [2004] have investigated the practical applicability of the Bell tree to coreference resolution. Each leaf node of the Bell tree represents a possible coreference outcome tree, and the problem of coreference resolution is cast as finding the best path from the root node to leaves. They use a maximum entropy model to rank paths in the Bell tree.

In Section 6 we compare those three machine learning approaches and conclude in the last section. Further terminology needed to understand the noun phrase coreference resolution is provided in next section.

2. Coreference Resolution

2.1. Basic Terminology

Natural languages provide speakers with a variety of ways to refer to entities. Two referring expressions that are used to refer to the same real-world entity are said to **corefer**. Reference to an entity that has been previously introduced into the discourse is called **anaphora**. Anaphor is the referring expression and the entity to which it refers is its **antecedent**. An anphora and all its antecedents form a coreference sequence called **coreferential chain**. A typical coreference resolution system (depicted in Figure 1) takes an arbitrary document as input and produces the appropriate coreferential chains as output.

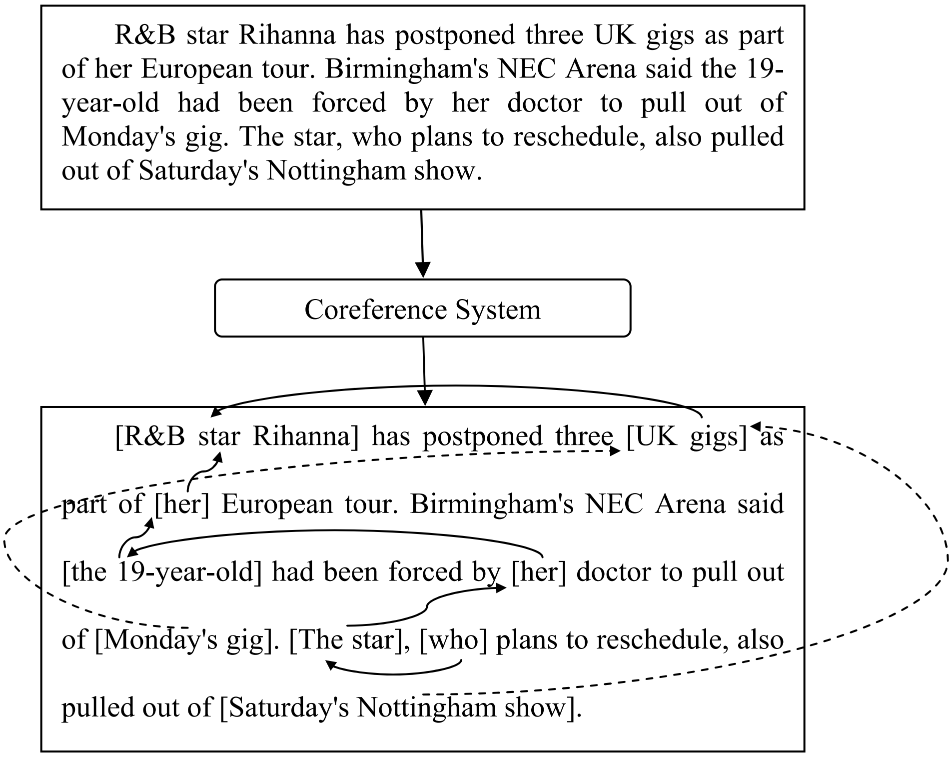

Figure 1. Coreference System: A full arrow represents anaphora; a dashed arrow stands for brigding anaphora, reference to the antecedent by generic knowledge (part-whole/element-set).

2.2. Evaluation

Precision and **recall** are two widely used measures for evaluating the quality of results. Precision can be seen as a measure of exactness, whereas recall is a measure of completeness. In a coreference resolution task, the precision is the number of noun phrase pairs correctly labeled as coreferent (true positives) divided by the total number of pairs labeled as coreferent (i.e. the sum of true positives and false positives, which are pairs incorrectly labeled as coreferent). Recall in this context is defined as the number of true positives divided by the total number of pairs that actually corefer (i.e. the sum of true positives and false negatives, which are pairs which were not labeled as coreferent but should have been).

Usually, precision and recall scores are combined into a single measure, the **F-measure**, which is the weighted harmonic mean of precision and recall.

	Correct classification	
Obtained classification	true positive (TP)	false positive (FP)
	false negative (FN)	true negative (TN)

Figure 2. Comparison between the given classification of a noun phrase pair and the desired correct classification.

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{\text{number of correctly predicted coreference links}}{\text{number of all predicted links}}$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{\text{number of correctly predicted coreference links}}{\text{number of all coreference links}}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3. Clustering Approach

Cardie and Wagstaff's [1999] unsupervised corpus-based clustering approach to the coreference task stems from the observation that each group of coreferent noun phrases defines an equivalence class. They start at the end of the document and compare each noun phrase to all preceding noun phrases. If the distance between two noun phrases is less than the clustering radius threshold r and their coreference equivalence classes are compatible, then the classes are merged.

The distance between two noun phrases is measured by a feature's weight and incompatibility function for each feature from the NP feature set. The NP feature set consists of word, head noun, position, pronoun type, article, words-substring, appositive, number, proper name, semantic class, gender and animacy. The incompatibility function returns a value between 0 and 1.

$$\text{dist}(NP_i, NP_j) = \sum_{f \in F} w_f \times \text{incompatibility}_f(NP_i, NP_j)$$

If two noun phrases do not match in number/proper names/class/gender/animacy feature, the distance between them gets a value of ∞ representing the incompatibility. Conversely, the appositive and words-substring terms with a weight of $-\infty$ force coreference with compatible values.

In an evaluation on the *MUC-6* [1995] coreference resolution corpus, Cardie and Wagstaff's clustering approach achieves the best F-measure of 53.6% with $r = 4$.

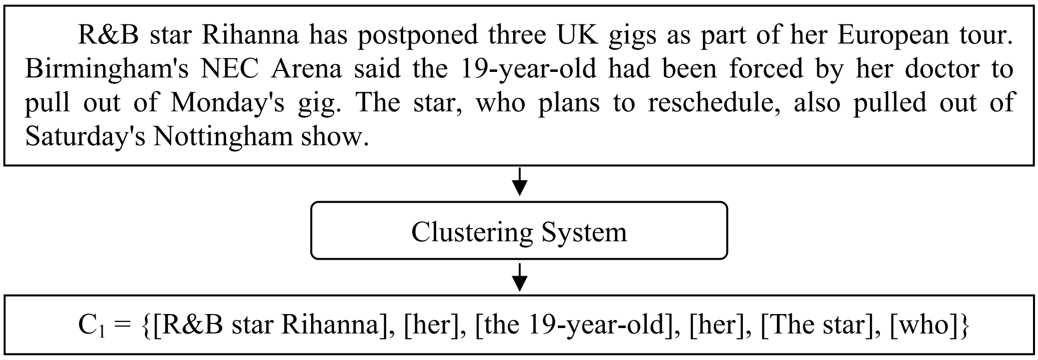

Figure 3. Coreference equivalence class in the sample text.

4. Decision Tree Algorithm

Decision tree algorithm uses a decision tree as a classifier model. In the tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. Applying decision tree algorithm for coreference resolution requires a set of features describing pairs of noun phrases and recasting the coreference problem as a classification task (e.g. [*Aone and Bennet*, 1995], [*McCarthy and Lehnert*, 1995], [*Soon et al.*, 2001]). A noun phrase coreference system described by Ng and Cardie [2002] extends the Soon et al. corpus-based approach.

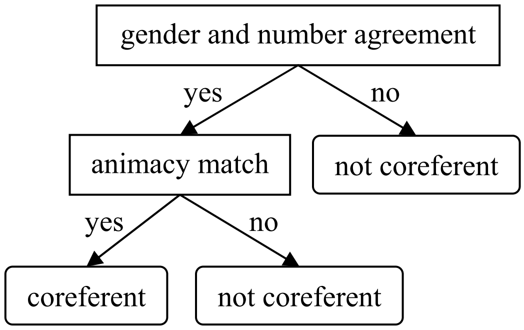

Figure 4. Simplified decision tree for coreference resolution.

Firstly, Ng and Cardie build a noun phrase coreference classifier using the C4.5 decision tree induction system. For a non-pronominal noun phrase, the closest non-pronominal preceding antecedent is selected to generate the positive training example. For pronouns, the closest preceding antecedent is selected. After training, texts are processed from left to right. Each noun phrase encountered is compared in turn to each preceding noun phrase from right to left. For each pair the coreference classifier returns a number between 0 and 1. Noun phrase pairs with class values above 0.5 are considered COREFERENT; otherwise the pair is considered NOT COREFERENT. The noun phrase with the highest coreference likelihood value from among preceding NPs with coreference class values above 0.5 is selected as the antecedent. The process terminates as soon as the antecedent is found or the beginning of the text is reached.

In the Ng and Cardie's coreference system a set of 53 features was proposed. The features were not derived empirically from the corpus, but were based on common-sense knowledge and linguistic intuitions regarding coreference. Surprisingly, the results using the full feature set are significantly low when compared with the results with a manual feature selection, with an eye toward eliminating low-precision rules for common noun resolution – F-measure of 70.4% on the MUC-6 coreference data sets and 63.4% on *MUC-7* [1998].

5. Algorithm Based on the Bell Tree

Luo et al. [2004] use the Bell tree to model the process of partitioning mentions into entities[1]. They traverse mentions in a document from beginning to end. The root node consists of a partial entity containing the first mention in the document. In each step of the algorithm, one mention is added by either linking to each of existing entities, or starting a new entity. A new layer of nodes is created to represent all possible coreference outcomes by adding one mention. The number of tree leaves is the number of possible coreference outcomes and it equals the Bell number[2] [*Bell*, 1934].

Since the Bell number increases rapidly as the number of mentions increases, pruning is necessary. Thus, instead of finding the best leaf node Luo et al. look for the best path from the root to leaves in the Bell tree. The algorithm uses maximum entropy model [*Berger et al.*, 1996] to rank paths and prunes any children with an insufficient score.

In the maximum entropy model a set of basic and composite features is selected. Composite features are generated by taking conjunction of basic features. Testing the algorithm on the MUC6 data Luo et al.'s system has 85.7% F-measure when using the official MUC scorer [Vilain et al., 1995].

---

[1] In the Luo et al.'s paper, a *mention* is defined as a referring expression, which can be all kinds of noun phrases, and the collection of mentions referring to the same object form an *entity* (by another name an equivalence class, used in the Cardie and Wagstaff's work).

[2] The Bell Number *B(n)* is the number of ways of partitioning *n* distinguishable objects (i.e., mentions) into non-empty disjoint subsets (i.e., entities).

$$B(n) = \frac{1}{e}\sum_{k=0}^{\infty} \frac{k^n}{k!}$$

6. Discussion

All three presented machine learning approaches created a set of features describing noun phrases and their relationship. In Cardie and Wagstaff's the set consists of 12 features. With the feature *article* (indefinite, definite, none) and *semantic class* (WordNet [Miller, 1990]) this clustering approach seems to be a language-dependent one which can only be used for English. Ng and Cardie offer three different string match features which restrict the application of string matching to pronouns, proper names, and non-pronominal NPs, respectively. Other Ng and Cardie's features to consider are MAXIMALNP (Do both NPs have the same maximal NP projection?), PREDNOM (Do the NPs form a predicate nominal construction?), SPAN (Does one NP span the other?), BINDING (Do the NPs violate conditions B or C of the Binding Theory?), CONTRAINDICES (Can NPs be co-indexed based on simple heuristics?). The interesting thing on Luo et al.' system is the total number of features which reaches $171K$; Most are conjunction features. In contrast to Ng and Cardie's work the full model with all $171K$ features gives also the best result.

7. Conclusion

In this paper we introduced three different machine learning approaches to coreference resolution. Each of them brings new useful information about features selection, training set creation, classifier building and the final noun phrases clustering. We believe that thanks to this knowledge the coreference resolution system for Czech [Nguy, 2006] [Nguy and Žabokrtský, 2007] can be further improved in the future.

Acknowledgments. This work has been supported by the MŠMT ČR LC536 grant.

References

C. Aone and S. W. Bennett, Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies, in: *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 122-129, 1995.

E. T. Bell, Exponential numbers, in: *Amer. Math. Monthly*, 411–419, 1934.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra, A maximum entropy approach to natural language processing, in: *Computational Linguistics*, 22(1):39–71, 1996.

A. Blum and T. Mitchell, Combining labeled and unlabeled data with Co-Training, in: *Proceedings of the 11th Annual Coreference on Learning Theory*, 92-200, 1998.

C. Cardie and K. Wagstaff, Noun phrase coreference as clustering, in: *Proceedings of the 1999 SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 82–89, 1999.

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos, A mention-synchronous coreference resolution algorithm based on the bell tree, in: *Proc. of ACL'04*, 2004.

J. McCarthy and W. Lehnert, Using Decision Trees for Coreference Resolution, in: *Proceedings of the Fourteenth International Conference on Artificial Intelligence*, 1050-1055, 1995.

G. A. Miller, WordNet: An on-line lexical database, in: *International Journal of Lexicography*, 3(4)~, 1990.

MUC-6, Proceedings of the Sixth Message Understanding Conference, Morgan Kaufmann, San Francisco, CA, 1995.

MUC-7, Proceedings of the Seventh Message Understanding Conference, Morgan Kaufmann, San Francisco, CA, 1998.

V. Ng and C. Cardie, Improving machine learning approaches to coreference resolution, in: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 104-111, 2002.

Nguy G. L., Proposal of a set of rules for anaphora resolution in Czech, Master's thesis, Faculty of Mathematics and Physics, Charles University, 2006.

Nguy G. L. and Z. Žabokrtský, Rule-based Person Pronominal Anaphora Resolution Method Using the Prague Dependency Treebank 2.0 Data, in: *Proceedings of DAARC 2007 (6th Discourse Anaphora and Anaphor Resolution Colloquium)*, 2007.

J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.

W. M. Soon, H. T. Ng and D. C. Y. Lim, A Machine Learning Approach to Coreference Resolution of Noun Phrases, in: *Computational Linguistics*, 27(4):521-544, 2001.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman, A model-theoretic coreference scoring scheme, in: *In Proc. of MUC6*, 45–52, 1995.