RELAXCOR: An Open Source Coreference Resolution System

Emili Sapena, Lluís Padró and Jordi Turmo*
TALP Research Center
Universitat Politècnica de Catalunya
Barcelona, Spain
{esapena, padro, turmo}@lsi.upc.edu

Abstract

RELAXCOR is a constraint-based graph partitioning approach to coreference resolution solved by relaxation labeling. It is an open source software, available for download, that can be used for research proposes in issues related to coreference resolution as well as a black box processing tool. This paper describes the architecture of the system and its advantages.

1 Introduction

Coreference resolution is a natural language processing task which consists of determining the *mentions* that refer to the same entity in a text or discourse. A mention is a noun phrase referring to an entity and includes named entities, definite noun phrases, and pronouns. For instance, *"Michael Jackson"* and *"the youngest of Jackson 5"* are two mentions referring to the same entity, i.e. those mentions corefer.

Many real world applications related to natural language relay on coreference resolution. Consider tasks such as machine translation (Peral et al., 1999), question answering (Morton, 2000) and summarization (Azzam et al., 1999). The higher their comprehension of a discourse, the better they perform.

RELAXCOR (Sapena et al., 2010a) is a constraint-based graph partitioning approach to coreference resolution solved by relaxation labeling. It is an open source software, available for download[1], that can help for the research of issues related to coreference resolution and can also be useful as a processing tool. The performances of RELAXCOR are in the state of the art, achieving the second position at CoNLL-2011 Shared Task (Pradhan et al., 2011).

There are already some coreference resolution systems publicly available such as GUITAR (Steinberger et al., 2007), BART (Versley et al., 2008), the Illinois Coreference Package (Bengtson and Roth, 2008), RECONCILE (Stoyanov et al., 2010), the Stanford's Multi-Pass Sieve Coreference Resolution System (Lee et al., 2011), and OPENNLP[2].

The main advantages of using RELAXCOR are the language adaption and the possibility to incorporate handwritten constraints, or constraints acquired from other sources. Regarding the languages, RELAXCOR is ready to work in English, Spanish and Catalan, and the incorporation of new languages requires minimal changes in the software. And regarding the incorporation of constraints, it can be useful for example for linguists, to introduce constraints by hand and evaluate their performances.

In addition, the software package includes the scorer used in Semeval-2010 task 1 (Recasens et al., 2010) and CoNLL-2011 Shared task.

2 System description

RELAXCOR is a coreference resolution system based on constraint satisfaction. It represents the

* Research supported by the Spanish Science and Innovation Ministry, via the KNOW2 project (TIN2009-14715-C04-04) and from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement number 247762 (FAUST)

[1] Privisional URL: http://www.lsi.upc.edu/~esapena/downloads/index.php?id=4
[2] http://opennlp.sourceforge.net

problem as a graph connecting any pair of candidate coreferent mentions and applies relaxation labeling, over a set of constraints, to decide the set of most compatible coreference relations. This approach combines classification and clustering in one step. Thus, decisions are taken considering the entire set of mentions, which ensures consistency and avoids local classification decisions. The RELAXCOR implementation is 90% Perl and 10% C++ and it is an improved version of the system that participated in the SemEval-2010 Task 1 (Sapena et al., 2010b) and CoNLL-2011 Shared task (Sapena et al., 2011).

RELAXCOR expects a preprocessed document in the input. The format of the input is a document where each row is a token and each column corresponds to a linguistic layer such as part of speech or parsing. The format is the same used in corpora Ontonotes (Pradhan et al., 2007) and AnCora-CO (Recasens and Martí, 2009). The order and formats of the columns can be configured. The minimum information required by the system is tokenization, part of speech, and dependency parsing, while named entities is not strictly necessary but really helpful. Other information like syntactic parsing, lemmatization, semantic role labeling and speaker is optional.

The resolution process of the system follows four steps:

1 Detect mentions.

2 Generate feature vectors for each pair of mentions.

3 Apply the set of constraints to all the pairs of mentions.

4 Solve coreferences using relaxation labeling over a graph.

2.1 Mention detection system

The mention detection system uses part of speech and syntactic information. Syntactic information may be dependency parsing or constituent parsing. The system extracts one candidate mention for every: Noun phrase (NP), pronoun, Named Entity, and capitalized common noun or proper name that appear two or more times in the document. In case that some NPs share the same head, the larger NP is selected and the rest discarded. Also the mention repetitions with exactly the same boundaries are discarded.

This mention detection system achieves an acceptable recall (p.e. higher than 90% in Ontonotes), but a low precision because includes many singletons. Note that a mention detection system in pipe configuration acts as a filter and the main objective at this point is to achieve as much recall as possible.

2.2 Features

The system has over a hundred features, binarized for each possible value. Each linguistic layer have a set of features that evaluate one mention of the pair or the compatibility of both in some criteria. For instance, lexical features include string comparisons, morphological features compare gender and number, syntactic features determine whether a mention is a demonstrative NP, and so on. The complete list of features can be found in (Sapena et al., 2011). In case that input data does not include preprocessing information of some linguistic layer, the features of this level can be excluded of the model.

2.3 Constraints

The knowledge of the system is represented as a set of weighted constraints. Each constraint has an associated weight reflecting its confidence. The sign of the weight indicates that a pair or a group of mentions corefer (positive) or not (negative). Only constraints over pairs of mentions are implemented in the current version of RELAXCOR.

The constraints are conjunctions of feature-value pairs. Moreover, given that features have been binarized, a constraint is just a conjunction of activated/negated features. Figure 1 is an example of a constraint.

The machine learning process generates a decision tree from the training data set and extracts a set of constraints with the C4.5 rule-learning algorithm (Quinlan, 1993). Then, constraints are applied to the training data in order find their weights. But constraints can be added from any source and can be manually written. Writing a constraint is as easy as writing in a text file the names of the features implied and the desired values. The weight of the constraint can be also determined by hand, but in case

DIST_SEN_1 & GENDER_YES & $\overline{I\_FIRST}$ &
I_MAXIMALNP & J_MAXIMALNP &
I_SRL_ARG_0 & J_SRL_ARG_0 &
I_TYPE_P & J_TYPE_P

Figure 1: Example of a constraint. It applies when the distance between m_i and m_j is exactly 1 sentence, their gender match, both are maximal NPs, both are argument 0 (subject) of their respective sentences, both are pronouns, and m_i is not the first mention of its sentence.

that having training/development data, the training process can be executed without learning constraints but learning the weights and other parameters.

2.4 Resolution

The coreference resolution problem is represented as a graph with mentions in the vertices. Mentions are connected to each other by edges. Edges are assigned a weight that indicates the confidence that the mention pair corefers or not. More specifically, an edge weight is the sum of the weights of the constraints that apply to that mention pair. The larger the edge weight in absolute terms, the more reliable.

RELAXCOR uses relaxation labeling for the resolution of the graph partitioning process, satisfying as many constraints as possible. Relaxation labeling is an iterative algorithm that performs function optimization based on local information. More information about the resolution process can be found in the original paper (Sapena et al., 2010a).

3 Adaption to other languages

The system is ready to solve coreferences in English, Spanish, and Catalan. And it can be easily adapted to other languages. A preprocess pipeline in the target language is needed, including a column with EAGLES standard for part of speech (Leech and Wilson, 1996). Otherwise, the system won't have information about gender and number unless a set of rules are handwritten in the corresponding functions of the code (concretely, Mention.pm and Features.pm modules). Moreover, the system have some resources for the original languages such as nicknames, typical (fe)male names and gentiles. The same resources for the target language are useful to achieve good performances, but not mandatory.

Measure	Recall	Precision	F_1
mention-based CEAF	53.51	53.51	53.51
entity-based CEAF	44.75	38.38	41.32
MUC	56.32	63.16	59.55
B^3	62.16	72.08	67.09
BLANC	69.50	73.07	71.10
(CEAFe+MUC+B^3)/3	-	-	**55.99**

Table 1: Official test results on CoNLL (Ontonotes, English)

Measure	Recall	Precision	F_1
mention-based CEAF	75.33	75.33	75.33
entity-based CEAF	84.66	78.66	81.55
MUC	59.74	68.08	63.64
B^3	84.67	77.17	80.75
(CEAFe+MUC+B^3)/3	-	-	**75.31**

Table 2: Test results on Spanish (AnCora-CO)

4 Results

The following tables show the scores obtained by RELAXCOR in English (Table 1) and Spanish (Table 2). RELAXCOR achieves the second position in CoNLL-2011 Shared task. Note that the differences in the results between English and the other languages is because the differences in the corpora which includes annotated singletons and causes a boost in mention-based measures CEAF and B^3. The MUC score is not affected by this difference. The package include models trained with respective corpora useful enough to solve any input data of unseen documents in the same language.

5 Scorer

The software package also includes an scorer to evaluate the outputs in case that gold annotation is available. The scorer has implemented the measures: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF (entity based and mention based) (Luo, 2005), and BLANC (Recasens and Hovy, 2011). Some of these measures were developed for an ideal scenario where the input includes the boundaries of the mentions (true mentions) and their behavior in a system mentions scenario was not contemplated. So, the mapping of the system mentions over the gold mentions includes some modifications inspired by (Cai and Strube, 2010). Moreover, the scorer also evaluates mention detection with precision and recall. This scorer have been used in Semeval-2010 and CoNLL-2011 for official evaluation.

References

S. Azzam, K. Humphreys, and R. Gaizauskas. 1999. Using coreference chains for text summarization. In *Proceedings of the Workshop on Coreference and its Applications*, pages 77–84. Association for Computational Linguistics.

A. Bagga and B. Baldwin. 1998. Algorithms for scoring coreference chains. *Proceedings of the Linguistic Coreference Workshop at LREC*, pages 563–566.

E. Bengtson and D. Roth. 2008. Understanding the value of features for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics.

Jie Cai and Michael Strube. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of SIGDIAL*, pages 28–36, University of Tokyo, Japan.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA, June. Association for Computational Linguistics.

G. Leech and A. Wilson. 1996. Recommendations for the syntactic annotation of corpora.

X. Luo. 2005. On coreference resolution performance metrics. *Proc. of HLT-EMNLP*, pages 25–32.

T.S. Morton. 2000. Using coreference in question answering. *NIST SPECIAL PUBLICATION SP*, pages 685–688.

J. Peral, M. Palomar, and A. Ferrández. 1999. Coreference-oriented interlingual slot structure & machine translation. In *Proceedings of the Workshop on Coreference and its Applications*, pages 69–76. Association for Computational Linguistics.

S.S. Pradhan, E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2007. OntoNotes: A unified relational semantic representation. *INTERNATIONAL JOURNAL OF SEMANTIC COMPUTING*, 1(4):405.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*, Portland, Oregon, June.

J.R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand Index for coreference evaluation. *Natural Language Engineering*, doi:10.1017/S135132491000029X.

M. Recasens and M.A. Martí. 2009. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8, Uppsala, Sweden, July. Association for Computational Linguistics.

Emili Sapena, Lluís Padró, and Jordi Turmo. 2010a. A Global Relaxation Labeling Approach to Coreference Resolution. In *Proceedings of 23rd International Conference on Computational Linguistics, COLING*, Beijing, China, August.

Emili Sapena, Lluís Padró, and Jordi Turmo. 2010b. RelaxCor: A Global Relaxation Labeling Approach to Coreference Resolution. In *Proceedings of the ACL Workshop on Semantic Evaluations (SemEval-2010)*, Uppsala, Sweden, July.

Emili Sapena, Lluís Padró, and Jordi Turmo. 2011. Relaxcor participation in conll shared task on coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 35–39, Portland, Oregon, USA, June. Association for Computational Linguistics.

Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, and Karel Jeek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management: an International Journal*, 43(6):1663–1680.

Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom. 2010. Coreference resolution with Reconcile. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010 Short Papers)*, pages 156–161, Uppsala, Sweden.

Y. Versley, A. Moschitti, M. Poesio, and X. Yang. 2008. Coreference systems based on kernels methods. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 961–968. Association for Computational Linguistics.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. *Proceedings of the 6th conference on Message understanding*, pages 45–52.