RelaxCor Participation in CoNLL Shared Task on Coreference Resolution

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

Abstract

This paper describes the participation of RELAXCOR in the CoNLL-2011 shared task: "Modeling Unrestricted Coreference in Ontonotes". RELAXCOR is a constraint-based graph partitioning approach to coreference resolution solved by relaxation labeling. The approach combines the strengths of groupwise classifiers and chain formation methods in one global method.

1 Introduction

The CoNLL-2011 shared task (Pradhan et al., 2011) is concerned with intra-document coreference resolution in English, using Ontonotes corpora. The core of the task is to identify which expressions (usually NPs) in a text refer to the same discourse entity.

This paper describes the participation of RELAXCOR and is organized as follows. Section 2 describes RELAXCOR, the system used in the task. Next, Section 3 describes the tuning needed by the system to adapt it to the task issues. The same section also analyzes the obtained results. Finally, Section 4 concludes the paper.

2 System description

RELAXCOR (Sapena et al., 2010a) is a coreference resolution system based on constraint satisfaction. It represents the 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 used in this task is an improved version of the system that participated in the SemEval-2010 Task 1 (Recasens et al., 2010).

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 were used in the current version of RELAXCOR. However, RELAXCOR can handle higher-order constraints. Constraints can be obtained from any source, including a training data set from which they can be manually or automatically acquired.

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 process. Relaxation labeling is an iterative algorithm that performs function optimization based on local information. It has been widely used to solve NLP problems. An array of probability values

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)

Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, pages 35–39,
Portland, Oregon, 23-24 June 2011.

is maintained for each vertex/mention. Each value corresponds to the probability that the mention belongs to a specific entity given all the possible entities in the document. During the resolution process, the probability arrays are updated according to the edge weights and probability arrays of the neighboring vertices. The larger the edge weight, the stronger the influence exerted by the neighboring probability array. The process stops when there are no more changes in the probability arrays or the maximum change does not exceed an *epsilon* parameter.

2.1 Attributes and Constraints

For the present study, all constraints were learned automatically using more than a hundred attributes over the mention pairs in the training sets. Usual attributes were used for each pair of mentions (m_i,m_j) –where $i < j$ following the order of the document–, like those in (Sapena et al., 2010b), but binarized for each possible value. In addition, a set of new mention attributes were included such as SAME_SPEAKER when both mentions have the same speaker[1] (Figures 1 and 2).

A decision tree was generated from the training data set, and a set of constraints was extracted with the C4.5 rule-learning algorithm (Quinlan, 1993). The so-learned constraints are conjunctions of attribute-value pairs. The weight associated with each constraint is the constraint precision minus a balance value, which is determined using the development set. Figure 3 is an example of a constraint.

2.2 Training data selection

Generating an example for each possible pair of mentions produces an unbalanced dataset where more than 99% of the examples are negative (not coreferent), even more considering that the mention detection system has a low precision (see Section 3.1). So, it generates large amounts of not coreferent mentions. In order to reduce the amount of negative pair examples, a clustering process is run using the positive examples as the centroids. For each positive example, only the negative examples with distance equal or less than a threshold d are included in the final training data. The distance is computed as the number of different attribute values

[1]This information is available in the column "speaker" of the corpora.

Distance and position:
Distance between m_i and m_j in sentences:
DIST_SEN_0: same sentence
DIST_SEN_1: consecutive sentences
DIST_SEN_L3: less than 3 sentences
Distance between m_i and m_j in phrases:
DIST_PHR_0, DIST_PHR_1, DIST_PHR_L3
Distance between m_i and m_j in mentions:
DIST_MEN_0, DIST_MEN_L3, DIST_MEN_L10
APPOSITIVE: One mention is in apposition with the other.
I/J_IN_QUOTES: $m_{i/j}$ is in quotes or inside a NP or a sentence in quotes.
I/J_FIRST: $m_{i/j}$ is the first mention in the sentence.
Lexical:
STR_MATCH: String matching of m_i and m_j
PRO_STR: Both are pronouns and their strings match
PN_STR: Both are proper names and their strings match
NONPRO_STR: String matching like in Soon et al. (2001) and mentions are not pronouns.
HEAD_MATCH: String matching of NP heads
TERM_MATCH: String matching of NP terms
I/J_HEAD_TERM: $m_{i/j}$ head matches with the term
Morphological:
The number of both mentions match: NUMBER_YES, NUMBER_NO, NUMBER_UN
The gender of both mentions match: GENDER_YES, GENDER_NO, GENDER_UN
Agreement: Gender and number of both mentions match: AGREEMENT_YES, AGREEMENT_NO, AGREEMENT_UN
Closest Agreement: m_i is the first agreement found looking backward from m_j: C_AGREEMENT_YES, C_AGREEMENT_NO, C_AGREEMENT_UN
I/J_THIRD_PERSON: $m_{i/j}$ is 3rd person
I/J_PROPER_NAME: $m_{i/j}$ is a proper name
I/J_NOUN: $m_{i/j}$ is a common noun
ANIMACY: Animacy of both mentions match (person, object)
I/J_REFLEXIVE: $m_{i/j}$ is a reflexive pronoun
I/J_POSSESSIVE: $m_{i/j}$ is a possessive pronoun
I/J_TYPE_P/E/N: $m_{i/j}$ is a pronoun (p), NE (e) or nominal (n)

Figure 1: Mention-pair attributes (1/2).

inside the feature vector. After some experiments over development data, the value of d was assigned to 5. Thus, the negative examples were discarded when they have more than five attribute values different than any positive example. So, in the end, 22.8% of the negative examples are discarded. Also, both positive and negative examples with distance zero (contradictions) are discarded.

2.3 Development process

The current version of RELAXCOR includes a parameter optimization process using the development data sets. The optimized parameters are *balance* and *pruning*. The former adjusts the constraint weights to improve the balance between precision and recall as shown in Figure 4; the latter limits the number of neighbors that a vertex can have. Limiting

Syntactic:
I/J_DEF_NP: $m_{i/j}$ is a definite NP.
I/J_DEM_NP: $m_{i/j}$ is a demonstrative NP.
I/J_INDEF_NP: $m_{i/j}$ is an indefinite NP.
NESTED: One mention is included in the other.
MAXIMALNP: Both mentions have the same NP parent or they are nested.
I/J_MAXIMALNP: $m_{i/j}$ is not included in any other NP.
I/J_EMBEDDED: $m_{i/j}$ is a noun and is not a maximal NP.
C_COMMANDS_IJ/JI: $m_{i/j}$ C-Commands $m_{j/i}$.
BINDING_POS: Condition A of binding theory.
BINDING_NEG: Conditions B and C of binding theory.
I/J_SRL_ARG_N/0/1/2/X/M/L/Z: Syntactic argument of $m_{i/j}$.
SAME_SRL_ARG: Both mentions are the same argument.
I/J_COORDINATE: $m_{i/j}$ is a coordinate NP
Semantic:
Semantic class of both mentions match (the same as (Soon et al., 2001))
SEMCLASS_YES, SEMCLASS_NO, SEMCLASS_UN
One mention is an alias of the other:
ALIAS_YES, ALIAS_NO, ALIAS_UN
I/J_PERSON: $m_{i/j}$ is a person.
I/J_ORGANIZATION: $m_{i/j}$ is an organization.
I/J_LOCATION: $m_{i/j}$ is a location.
SRL_SAMEVERB: Both mentions have a semantic role for the same verb.
SRL_SAME_ROLE: The same semantic role.
SAME_SPEAKER: The same speaker for both mentions.

Figure 2: Mention-pair attributes (2/2).

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**
Precision: 0.9581
Training examples: 501

Figure 3: 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. The final weight will be $weight = precision - balance$.

the number of neighbors reduces the computational cost significantly and improves overall performance too. Optimizing this parameter depends on properties like document size and the quality of the information given by the constraints.

The development process calculates a grid given the possible values of both parameters: from 0 to 1 for balance with a step of 0.05, and from 2 to 14 for pruning with a step of 2. Both parameters were empirically adjusted on the development set for the evaluation measure used in this shared task: the unweighted average of MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998) and entity-based CEAF (Luo, 2005).

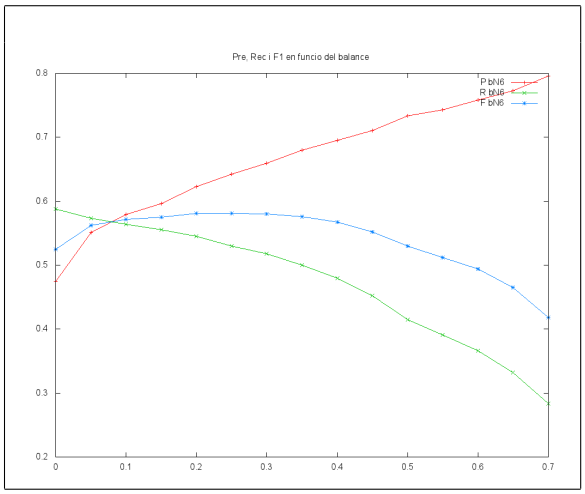

Figure 4: Development process. The figure shows MUC's precision (red), recall (green), and F_1 (blue) for each balance value with pruning adjusted to 6.

3 CoNLL shared task participation

RELAXCOR has participated in the CoNLL task in the Closed mode. All the knowledge required by the feature functions is obtained from the annotations of the corpora and no external resources have been used with the exception of WordNet (Miller, 1995), gender and number information (Bergsma and Lin, 2006) and sense inventories. All of them are allowed by the task organization and available in their website.

There are many remarkable features that make this task different and more difficult but realistic than previous ones. About mention annotation, it is important to emphasize that singletons are not annotated, mentions must be detected by the system and the mapping between system and true mentions is limited to exact matching of boundaries. Moreover, some verbs have been annotated as corefering mentions. Regarding the evaluation, the scorer uses the modification of (Cai and Strube, 2010), unprecedented so far, and the corpora was published very recently and there are no published results yet to use as reference. Finally, all the preprocessed information is automatic for the test dataset, carring out some noisy errors which is a handicap from the point of view of machine learning.

Following there is a description of the mention detection system developed for the task and an analysis of the obtained results in the development dataset.

3.1 Mention detection system

The mention detection system extracts one mention for every NP found in the syntactic tree, one for every pronoun and one for every named entity. Then, the head of every NP is determined using part-of-speech tags and a set of rules from (Collins, 1999). 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. In addition, nouns with capital letters and proper names not included yet, that appear two or more times in the document, are also included. For instance, the NP "*an Internet business*" is added as a mention, but also "*Internet*" itself is added in the case that the word is found once again in the document.

As a result, taking into account that just exact boundary matching is accepted, the mention detection achieves an acceptable recall, higher than 90%, but a low precision (see Table 1). The most typical error made by the system is to include extracted NPs that are not referential (e.g., predicative and appositive phrases) and mentions with incorrect boundaries. The incorrect boundaries are mainly due to errors in the predicted syntactic column and some mention annotation discrepancies. Furthermore, verbs are not detected by this algorithm, so most of the missing mentions are verbs.

3.2 Results analysis

The results obtained by RELAXCOR can be found in Tables 1 and 2. Due to the lack of annotated singletons, mention-based metrics B^3 and CEAF produce lower scores –near 60% and 50% respectively– than the ones typically achieved with different annotations and mapping policies –usually near 80% and 70%. Moreover, the requirement that systems use automatic preprocessing and do their own mention detection increase the difficulty of the task which obviously decreases the scores in general.

The measure which remains more stable on its scores is MUC given that it is link-based and not takes singletons into account anyway. Thus, it is the only one comparable with the state of the art right now. The results obtained with MUC scorer show an improvement of RELAXCOR's recall, a feature that needed improvement given the previous published results with a MUCs recall remarkably low (Sapena et al., 2010b).

Measure	Recall	Precision	F_1
Mention detection	92.45	27.34	42.20
mention-based CEAF	55.27	55.27	55.27
entity-based CEAF	47.20	40.01	43.31
MUC	54.53	62.25	58.13
B^3	63.72	73.83	68.40
(CEAFe+MUC+B^3)/3	-	-	56.61

Table 1: Results on the development data set

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 2: Official test results

4 Conclusion

The participation of RELAXCOR to the CoNLL shared task has been useful to evaluate the system using data never seen before in a totally automatic context: predicted preprocessing and system mentions. Many published systems typically use the same data sets (ACE and MUC) and it is easy to unintentionally adapt the system to the corpora and not just to the problem. This kind of tasks favor comparisons between systems with the same framework and initial conditions.

The obtained performances confirm the robustness of RELAXCOR and a recall improvement. And the overall performance seems considerably good taking into account the unprecedented scenario. However, a deeper error analysis is needed, specially in the mention detection system with a low precision and the training data selection process which may be discarding positive examples that could help improving recall.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement number 247762 (FAUST), and from the Spanish Science and Innovation Ministry, via the KNOW2 project (TIN2009-14715-C04-04).

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at LREC 98*, pages 563–566, Granada, Spain.

S. Bergsma and D. Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 33–40. 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.

M. Collins. 1999. *Head-driven statistical models for natural language parsing*. Ph.D. thesis, University of Pennsylvania.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP 2005*, pages 25–32, Vancouver, B.C., Canada.

G.A. Miller. 1995. WordNet: a lexical database for English.

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, 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.

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

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 45–52.