Rule and Tree Ensembles for Unrestricted Coreference Resolution[*]

Cicero Nogueira dos Santos
Universidade de Fortaleza – UNIFOR
Informática Aplicada – PPGIA
Fortaleza, Brazil
cnogueira@unifor.br

Davi Lopes Carvalho
Universidade de Fortaleza – UNIFOR
Informática Aplicada – PPGIA
Fortaleza, Brazil
davi.carvalho@gmail.com

Abstract

In this paper, we describe a machine learning system based on rule and tree ensembles for unrestricted coreference resolution. We use Entropy Guided Transformation Learning (ETL) and Decision Trees as the base learners, and, respectively, ETL Committee and Random Forest as ensemble algorithms. Our system is evaluated on the closed track of the CoNLL 2011 shared task: Modeling Unrestricted Coreference in OntoNotes. A preliminary version of our system achieves the 6th best score out of 21 competitors in the CoNLL 2011 shared task. Here, we depict the system architecture and our experimental results and findings.

1 Introduction

Unrestricted coreference resolution consists in identifying coreferring entities and events in texts. For instance, in the sentence

> "She had *a good suggestion* and
> *it* was unanimously accepted."

there is a coreference between the pronoun "*it*" and the noun phrase "*a good suggestion*". In the following sentence

> "Sales of passenger cars *grew* 22%. *The strong growth* followed year-to-year increases."

there is a coreference between the noun phrase "*the strong growth*" and the event "*grew*". Throughout this paper, we use the term *mention* to mean a reference to an entity or event.

The CoNLL 2011 Shared Task (Pradhan et al., 2011) is dedicated to modeling unrestricted coreference in OntoNotes. The participants are provided with a large corpus that contains various annotation layers such as part-of-speech (POS) tagging, parsing, named entities and semantic role labeling. The task consists in the automatic identification of coreferring entities and events given predicted information on other OntoNotes layers. A previous work on modeling unrestricted coreference using an earlier version of this corpus is presented in (Pradhan et al., 2007).

In this paper, we describe the machine learning approach that we used to the closed track of the CoNLL 2011 Shared Task. Our system follows the common strategy of recasting the problem as a classification task. First, in a preprocessing step, a set of candidate mentions is constructed. Next, also in the preprocessing step, pairs of candidate co-referring mentions are generated. Then, each candidate pair of mentions is classified as co-referring or not using a classifier learned from the annotated corpus. Finally, a postprocessing step (clustering) removes inconsistencies that would result of the pairwise classifications and constructs a partition on the set of mentions. In our system, the learning module is based on ensemble learning. We use Entropy Guided Transformation Learning (ETL) (Milidiú et al., 2008) and Decision Trees (DT) (Quinlan, 1993) as base learners, and, respectively, ETL Committee (dos Santos et al., 2010) and Random Forest (Breiman, 2001) as ensemble algorithms.

The remainder of this paper is organized as follows. In Section 2, we present the corpus preprocessing and postprocessing steps. Our machine learning modeling for the unrestricted coreference resolution task is presented in Section 3. The experimental findings are depicted in Section 4. Finally, in Section 5, we present our final remarks.

2 Corpus Processing

In this section we describe some preprocessing and postprocessing steps used in the proposed system.

2.1 Candidate Mention Extraction

For each text document, we generate a list of candidate mentions in the following way:

- all the noun phrases (NP) identified in the provided parsing tree are considered as candidate mentions;

- each pronoun is isolatedly considered as a candidate mention even if it is inside a larger NP;

- named entities in the categories Person (PERSON), Organization (ORG) and Geo-Political Entity (GPE) are isolatedly considered as candidate mentions even if they are inside larger NPs. Additionally, in order to better align with the OntoNotes mention annotation, a processing is performed to include possessive marks "'s" and premodifiers such as "Mr.".

In the current version, our system does not consider verbs when creating candidate mentions. Therefore, the system does not resolve coreferences involving events.

2.2 Candidate Co-referring Pairs Generation

In the training phase, we generate positive and negative examples of co-referring pairs using a strategy similar to the one of Soon et al. (2001). In their method, the text is examined in a left-to-right manner. For each anaphoric mention m_j, is generated a positive example pair that includes m_j and its closest preceding antecedent, m_i. A negative example is created for m_j paired with each of the intervening mentions, $m_{i+1}, m_{i+2}, ..., m_{j-1}$. We extend the Soon et al. (2001) approach by also including all positive and negative pairs that can be formed with

the mentions in the sentence of the closest preceding antecedent, m_i.

In the classification phase, the text is also examined in a left-to-right manner. For each mention m_j, candidate co-referring pairs are generated by pairing it with a limited number of preceding mentions. When using predicted mentions, we set this limit to 60 (sixty). For the gold-mentions track, the limit is set to 40 (forty).

2.3 Feature Engineering

We use a set of 80 features to describe each pair of mentions (m_i, m_j). The feature set includes lexical, morphological, syntactic, semantic and positional information. Most of them are borrowed from the works of Ng and Cardie (2002) and Sapena et al. (2010). However, we also propose some new features. In the following, due to space constraints, we briefly describe some of them. The features marked with * are the new proposed ones.

Lexical: *head word* of $m_{i/j}$; *String matching* of (head word of) m_i and m_j (y/n); *Both are pronouns* and their strings match (y/n); *Previous/Next two words* of $m_{i/j}$; *Length* of $m_{i/j}$; *Edit distance* of head words; $m_{i/j}$ is a definitive NP (y/n); $m_{i/j}$ is a demonstrative NP (y/n).

Morphological: *Both are proper names* and their strings match (y/n); *Basic gender agreement**, which use a list of proper names extracted from the training corpus (y/n); *Gender/Number* of $m_{i/j}$; *Gender/Number agreement*(y/n), this and the previous feature are generated using the number and gender data provided by Bergsma and Lin (2006).

Syntactic: *POS tag* of the $m_{i/j}$ head word; *Previous/Next two POS tags* of $m_{i/j}$; m_i and m_j are *both pronouns / proper names* (y/n); *Previous/Next predicate* of $m_{i/j}$*; *Compatible pronouns*, which checks whether two pronouns agree in number, gender and person (y/n)*; *NP embedding level*; *Number of embedded NPs* in $m_{i/j}$*.

Semantic: the result of a *baseline system*; *sense* of the $m_{i/j}$ head word; *Named entity* type of $m_{i/j}$; m_i and m_j have the *same named entity*; *Semantic role of $m_{i/j}$* for the prev/next predicate*; *Concatenation of semantic roles of m_i and m_j* for the same predicate (if they are in the same sentence)*; *Same speaker** (y/n); *Alias* (y/n); m_i and m_j have a *hypernym/hyponym* relation (y/n); m_i and m_j have the

same semantic class (y/n); *sum of distances* between m_i and m_j to their class. The last three features are generated using WordNet 3.0 (Miller, 1995).

Distance and Position: Distance between m_i and m_j in sentences; Distance in number of mentions; Distance in number of person names (applies only for the cases where m_i and m_j are both pronouns or one of them is a person name)*; One mention is in apposition to the other (y/n).

2.4 Clustering Strategy

In order to generate the coreference chains, it is needed a strategy to create a partition in the mentions using the predictions for the candidate coreferent pairs. This part of the coreference resolution system is frequently called *clustering strategy* (Ng and Cardie, 2002). Our system uses an aggressive-merge clustering approach similar to the one proposed by Mccarthy and Lehnert (1995). In this strategy, each mention is merged with all of its preceding mentions that are classified as coreferent with it.

Additionally, a postprocessing step is employed to remove inconsistencies that would result of the clustering processing, such as an NP being coreferent to its embedded NP.

3 Machine Learning Modeling

In this section we briefly describes the machine learning approaches used in our experiments. We also describe a baseline system (BLS) that is used by ETL for the learning of correction rules. The classification produced by the BLS is also used as a feature for the other experimented learning strategies.

ETL: Entropy Guided Transformation Learning (ETL) is a correction rule learning algorithm. It extends Transformation Based Learning (TBL) by automatically generating rule templates using Decision Trees (DT) (Milidiú et al., 2008). We use an in-house implementation of ETL.

ETL Committee: is an ensemble method that uses ETL as the base learner (dos Santos et al., 2010). This approach combines the main ideas of Bagging and Random Subspaces, as well as rule redundancy and template sampling to generate diverse ETL classifiers. We use an in-house implementation of ETL Committee.

Decision Trees: the C4.5 (Quinlan, 1993) system is one of the most popular DT induction implementation. It induces a tree based classifier using the training data information gain. In our experiments, we use the J48 tool, which is a DT induction system similar to C4.5. J48 is part of the WEKA data mining toolkit (Hall et al., 2009).

Random Forest: is an ensemble method that uses DT as the base learner. In the Random Forest learning process (Breiman, 2001), first, bootstrap sampling is employed to generate multiple replicates of the training set. Then, a decision tree is grown for each training set replicate. When growing a tree, a subset of the available features is randomly selected at each node, the best split available within those features is selected for that node. In our experiments, the WEKA's Random Forest implementation is used.

Baseline System: the BLS classifies a candidate co-referring pair (m_i, m_j) as co-referring when one of the following conditions occur:

- m_j is an alias of m_i;

- m_j and m_i are 3rd person pronouns and there is no person name between them;

- the pair is composed of a person name and a 3rd person pronoun and there is no person name between them;

- removing determiners, m_i matches m_j;

- the feature *basic gender agreement* is true.

The parameters of each algorithm are tuned using the development set. For both, ETL Committee and Random Forest the ensemble size is set to 50.

4 Experiments and Results

We train models for two different CoNLL 2011 shared task closed tracks: (a) using candidate mentions whose boundaries are automatically extracted (see Section 2.1); and (b) using candidate mentions whose boundaries are provided. In the training phase, the gold standard OntoNotes annotation layers are used. For the development and test sets the automatically generated OntoNotes annotation layers are used.

For all experiments, results are reported using three metrics: MUC, B^3 and CEAF(E). We also report the average F_1 score for these three metrics, which is the official CoNLL 2011 shared task metric. Additionally, results for the test set are also reported using the CEAF(M) and BLANC metrics.

4.1 Automatic Mention Boundaries

In Table 1, we show machine learning system results for unrestricted coreference resolution using the development set. As we can see in Table 1, the results of ensemble methods are better than ones of the base learners, which is the expected result. ETL Committee is the classifier that achieve the best results, closely followed by Random Forest.

All the experimented ML systems achieve results better than the baseline. However, the improvement provided by ML is more expressive only for the MUC metric. For instance, ETL Committee provides an improvement over the baseline of about 6.5 points in the MUC F1-score, while the improvement for the other two metrics is only about 2 points.

We run an additional experiment by constructing a heterogeneous committee composed by the three best classifiers: (1) ETL Committee, (2) Random Forest and (3) ETL. The results for this system is shown in table line with ML Model name "(1) + (2) + (3)". This heterogeneous committee provides our best experimental results for the development set, which is slightly better than ETL Committee results.

Due to deadline constraints, the system output that we have submitted to the CoNLL 2011 shared task is a majority voting committee of three different ETL classifiers. These three ETL classifiers slightly differs in the used feature sets. In Table 1, the results of the Submitted System is presented for the development set. Table 2 presents the Submitted System results for the test set. Our system achieves the 6th best score out of 21 competitors in the closed track of the CoNLL 2011 shared task.

4.2 Gold Mention Boundaries

For the gold mention boundaries task, we were not able to assess system performances on the development set. This is due to the fact that not all gold mentions are annotated in the development set.

We have submitted two outputs for the CoNLL 2011 shared task gold mentions closed track. These

Metric	R	P	F_1
MUC	59.21	54.30	56.65
BCUBED	68.79	62.81	65.66
CEAF (M)	49.54	49.54	49.54
CEAF (E)	35.86	40.21	37.91
BLANC	73.37	66.91	69.46
(MUC + B^3 + CEAF(E))/3			53.41

Table 2: Submitted System results for the test set using automatically extracted mention boundaries.

outputs were generated by two systems described in the previous subsection: (a) the Submitted System; and (b) the heterogeneous committee (ETL Committee + Random Forest + ETL). In Table 3, we show the system results for the test set with gold standard mentions. Again, the heterogeneous committee provides our best results.

At the moment of writing this paper, the scoreboard for this task has not yet been released by the CoNLL 2011 shared task committee.

5 Conclusion

In this paper, we describe a machine learning system based on rule and tree ensembles for unrestricted coreference resolution. The system uses Entropy Guided Transformation Learning and Decision Trees as the base learners. ETL Committee and Random Forest are the used ensemble algorithms. We depict the system architecture and present experimental results and findings of our participation in the CoNLL 2011 shared task.

We present results for two closed tasks: (a) using automatically extracted mention boundaries; and (b) using gold mention boundaries. For both tasks, ensemble classifiers have better results than the base classifiers. This is the expected outcome, since ensemble classifiers tend to be more accurate than base classifiers. We also experiment heterogeneous committees that combines the three best classifier for the first task. Heterogeneous committees provide our best scoring results for both tasks. Using a preliminary version of our system, we achieve the 6th best score out of 21 competitors in the closed track of the CoNLL 2011 shared task.

One of the possible future works, is to investigate the impact of the new features that we propose.

ML Model	MUC			B³			CEAF(E)			(MUC + B³ + CEAF(E))/3
	R	P	F₁	R	P	F₁	R	P	F₁	F₁
(1) ETL Committee	52.31	**57.51**	54.78	63.62	70.42	66.84	42.64	37.99	40.18	53.93
(2) Random Forest	53.31	54.91	54.10	65.23	67.31	66.25	40.47	39.05	39.75	53.37
(3) ETL	**54.80**	52.24	53.49	67.56	62.19	64.77	37.22	39.55	38.35	52.20
(4) Decision Trees	57.51	49.12	52.98	**71.23**	58.94	64.50	34.84	**42.25**	38.19	51.89
(5) Baseline System	43.04	55.13	48.34	57.82	**74.21**	64.99	**43.63**	33.62	37.98	50.43
(1) + (2) + (3)	52.77	57.44	**55.00**	64.09	70.58	**67.18**	42.67	38.48	**40.47**	**54.21**
Submitted System	54.65	53.25	53.94	67.15	63.86	65.46	38.3	39.56	38.92	52.45

Table 1: System results for the development set using automatically extracted mention boundaries.

Metric	Submitted System			(1) + (2) + (3)		
	R	P	F₁	R	P	F₁
MUC	**58.77**	56.54	57.64	57.76	**61.39**	**59.52**
BCUBED	**67.05**	64.84	65.92	64.49	**70.27**	**67.26**
CEAF (M)	50.05	50.05	50.05	**51.87**	**51.87**	**51.87**
CEAF (E)	37.61	**39.62**	38.59	**41.42**	38.16	**39.72**
BLANC	72.59	67.76	69.80	**72.72**	**71.97**	**72.34**
(MUC + B³ + CEAF(E))/3			54.05			**55.50**

Table 3: System results for the test set using gold mention boundaries.

References

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of ACL2006*, ACL–44, pages 33–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Cícero Nogueira dos Santos, Ruy Luiz Milidiú, Carlos E. M. Crestana, and Eraldo R. Fernandes. 2010. ETL ensembles for chunking, NER and SRL. In *11th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing*, pages 100–112.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *Sigkdd Explorations*, 11:10–18.

Joseph F Mccarthy and Wendy G Lehnert. 1995. Using decision trees for coreference resolution. In *In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1050–1055.

Ruy L. Milidiú, Cícero N. dos Santos, and Julio C. Duarte. 2008. Phrase chunking using entropy guided transformation learning. In *Proceedings of ACL2008*, Columbus, Ohio.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 104–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted Coreference: Identifying Entities and Events in OntoNotes. In *in Proceedings of the IEEE International Conference on Semantic Computing (ICSC)*, September 17–19.

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. Ross Quinlan. 1993. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Emili Sapena, Lluís Padró, and Jordi Turmo. 2010. Relaxcor: A global relaxation labeling approach to coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 88–91, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27:521–544, December.