Corry: A System for Coreference Resolution

Olga Uryupina
CiMeC, University of Trento
uryupina@gmail.com

Abstract

Corry is a system for coreference resolution in English. It supports both local (Soon et al. (2001)-style) and global (Integer Linear Programming, Denis and Baldridge (2007)-style) models of coreference. Corry relies on a rich linguistically motivated feature set, which has, however, been manually reduced to 64 features for efficiency reasons. Three runs have been submitted for the SemEval task 1 on Coreference Resolution (Recasens et al., 2010), optimizing Corry's performance for BLANC (Recasens and Hovy, in prep), MUC (Vilain et al., 1995) and CEAF (Luo, 2005). Corry runs have shown the best performance level among all the systems in their track for the corresponding metric.

1 Introduction

Corry is a system for coreference resolution in English. It supports both local (Soon et al. (2001)-style) and global (ILP, Denis and Baldridge (2007)-style) models of coreference. The backbone of the system is a family of SVM classifiers for pairs of mentions: each mention type receives its own classifier. A separate anaphoricity classifier is learned for the ILP setting. Corry relies on a rich linguistically motivated feature set, which has, however, been manually reduced to 64 features for efficiency reasons.

Corry has only participated in the "open" setting, as it has already a number of preprocessing modules integrated into the system: the Stanford NLP toolkit for parsing (Klein and Manning, 2003) and NE-tagging (Finkel et al., 2005), Wordnet for semantic classes and the U.S. census data for assigning gender values to person names.

Three runs have been submitted for the SemEval task 1 on Coreference Resolution, optimizing Corry's performance for BLANC, MUC and CEAF. The runs differ with respect to the model (local for BLANC, global for MUC and CEAF) and the definition of mention types.

2 Preprocessing and Mention Extraction

In our previous study (Uryupina, 2008) we have shown that up to 35% recall and 20% precision errors in coreference resolution for MUC corpora are due to inaccurate mention detection. We have therefore invested substantial efforts into our mention detection module.

Most state-of-the-art coreference resolution systems operate either on *gold* markables or on the output of an ACE-style mention detection module. We are not aware of extensive studies on mention extraction algorithms for such datasets as SemEval (OntoNotes) where mentions are complex NPs not constrained with respect to their semantic types.

We rely on the Stanford NLP toolkit for extracting named entities (Finkel et al., 2005) and parse trees for each sentence (Klein and Manning, 2003). We then merge the output of the NE-tagger and the parser to create a list of mentions in the following way:

1. Named entities are considered mentions if they correspond to a sequence of parsing constraints.
2. Pronouns are considered mentions if they are not a part of an NE-mention.
3. NPs are considered "candidate mentions" if they are not a part of an NE-mention. The set of

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 100–103,
Uppsala, Sweden, 15-16 July 2010.

candidate mentions is then filtered to eliminate pairs of NPs with the same head noun (coordinate NPs receive unique artificial heads). For possessive NPs we adjust the boundaries and the head to exclude the "'s" token. The remaining candidates are aligned with NE-mentions – if an NE and an NP have the same last word, they are considered the same mention of a special type. Finally, the list of candidates is optionally filtered using a small stop-list (for example, all the "there" NPs in "There is .." are discarded).

We rely on the Stanford NLP toolkit, WordNet and the U.S. census data to assign numerous properties to our mentions: semantic type, number, gender and others.

3 Features

Corry relies on two SVM[1] classifiers for *coreference* and *anaphoricity*. The former determines whether two given mentions M_i and M_j are coreferent or not. The latter determines whether a given mention M_i is anaphoric or discourse new. In Section 4 we show how these classifiers help us build coreference chains. We use the SVM-Light package (Joachims, 1999) for learning our classifiers.

The strength of our system lies in its rich feature set for the coreference classifier. In our previous studies (Uryupina, 2006; 2007) we have tested up to 351 nominal/continuous (1096 boolean/continuous) features showing significant improvements over basic feature sets advocated in the literature. For the SemEval task 1, we have reduced our rich feature set to 64 nominal/continuous features for efficiency reasons: on the one hand, our new set is large enough to cover complex linguistic patterns of coreference, on the other hand, it allows us to test different settings and investigate possibilities for global modeling.

Our *anaphoricity* classifier is used by the ILP model. It relies on 26 boolean/continuous features. More details on the classifier itself can be found in (Uryupina, 2003).

[1]Corry supports a number of machine learning algorithms: C4.5, TiMBL, Ripper, MaxEnt and SVM. See Uryupina (2006) for a comparison of Corry's performance with different learners.

4 Modeling

Corry supports both global and local views of coreference. Our evaluation experiments (cf. Section 5) show that the choice of a particular model should be motivated by the desired scoring metric.

Our local model of coreference is a reimplementation of the algorithm, proposed by Soon et al. (2001) with an extended feature set. The core of Soon et al.'s (2001) approach is a *link*-based classifier: it determines whether a given pair of markables are coreferent or not. During testing, a greedy clustering algorithm (link-first) is next used to build coreference chains on the output of the classifier.

We have slightly extended this model to allow separate classifiers for different *mention types*: each candidate anaphor receives a type (e.g. "pronoun") and is processed with a corresponding classifier. We, thus, rely on a family of classifiers, with the same feature set and the same machine learner. The exact definition of mention types is a parameter to be determined empirically on the development set.

Our global model is largely motivated by Denis and Baldridge (2007; 2008) and Finkel and Manning (2008). Following these studies, we use Integer Linear Programming to find the most globally optimal solution, given the decisions made by our *coreference* and *anaphoricity* classifiers.

In general, an ILP problem is determined by an objective function to be maximized (or minimized) and a set of task-specific constraints. The function is defined by costs $link_{<i,j>}$, and $dnew_j$ reflecting potential gains and losses for committing to specific variable assignments. We assume that costs can be positive (for pairs of markables that are likely to be coreferent) or negative (for pairs of markables that are unlikely to be coreferent). The costs are computed by an external module (such as a family of local classifiers described above). The objective function then takes the form:

$$max\Big(\sum_{<i,j>} link_{<i,j>} * L_{<i,j>} - \sum_{j} dnew_j * D_j\Big) \quad (1)$$

Binary variables $L_{<i,j>}$ indicate that two markables M_i and M_j are coreferent in the output assignment. Binary variables D_j indicate that the markable M_j is considered anaphoric in the output assignment. The ILP solver thus assigns values to

$L_{<i,j>}, \forall i,j : i < j$ and $D_j, \forall j$ whilst maximizing the objective in (1). We take the transitive closure of all the proposed $L_{<i,j>}$ to build the output partition.

Note that the objective in (1) is not constrained in any way and will thus allow illegal variable assignments. For example it does not constrain the assignment of L and D variables to be consistent with one another and does not enforce transitivity. The following constraints suggested in the literature (Denis and Baldridge, 2007; Denis and Baldridge, 2008; Finkel and Manning, 2008) ensure that these and other coreference properties are respected:

1. Best-link constraint

$$B : \sum_i L_{<i,j>} \leq 1, \forall j \quad (2)$$

2. Transitivity constraints

$$\forall i,j,k : i < j < k$$
$$T : \quad L_{<i,j>} + L_{<j,k>} - 1 \leq L_{<i,k>} \quad (3)$$
$$L : \quad L_{<j,k>} + L_{<i,k>} - 1 \leq L_{<i,j>} \quad (4)$$
$$R : \quad L_{<i,j>} + L_{<i,k>} - 1 \leq L_{<j,k>} \quad (5)$$

3. Anaphoricity constraints

$$A : \quad \textstyle\sum_i L_{<i,j>} >= D_j \; \forall j \quad (6)$$
$$D : \quad L_{<i,j>} \leq D_j \; \forall i,j \quad (7)$$

We refer the reader to the above-mentioned papers for detailed discussions of these constraints and their impact on coreference resolution. As we show in Section 5 below, the usability of a particular constraint should be determined experimentally based on the desired system behaviour.

5 Evaluation

5.1 Development

Corry has participated in the *gold* and *regular* open settings for English. We have collected a number of runs on the development data to optimize the performance level for a particular score: BLANC (Recasens and Hovy, in prep), MUC (Vilain et al., 1995) or CEAF (Luo, 2005). The runs differ with respect to the model (local vs. global with varying sets of constraints) and the definition of mention types. We deliberately left the B-CUBE score (Bagga and Baldwin, 1998) completely out of our preliminary experiments. The official SemEval scorer was used for these experiments.

Our experiments on the development set show that no configuration is able to produce equally reliable scores according to all the metrics (note, for example, that on the test set the BLANC difference between Corry-M and Corry-B in the *gold* setting is almost 10%). We believe that it is a challenging point for future research.

We have selected the best configurations for each score and submitted them as separate runs. The Corry-C system, optimized for CEAF-$\phi 4$, is a global model with the L, D and A constraints. For the *gold* setting, mention types are defined as pronouns and non-pronouns. For the *regular* setting, the system distinguishes between "speech" pronouns, 3rd person pronouns, names and nominals.

Corry-M, optimized for MUC, is a global model with the D constraint and separate classifiers for pronouns, names and nominals. Note that, compared to Corry-C, this setting allows for more coreference links – it is well known from the literature (cf., for example, Bagga and Baldwin (1998)) that the MUC metric is biased towards recall.

Finally, Corry-B, optimized for BLANC, is a local model that distinguishes between pronouns, nominals and names. The fact that such a simple model is able to outperform much more complex versions of Corry strengthens the importance of feature engineering.

5.2 Testing

Table 1 shows the SemEval task 1 scores for the gold/regular open setting. Corry has shown reliable performance for both mention detection and coreference resolution. For mention detection, Corry's F-score is 4% higher than the one of the competing approach. For coreference, all the Corry runs yielded the best performance level for a score under optimization.

Finally, for the B-CUBE metric that had not been optimized at all, Corry lost only marginally to the RelaxCor system in the gold setting and came first in the regular setting.

6 Conclusion

We have presented Corry – a system for coreference resolution in English. Our plans include extending it to cover multiple languages. However, as the main strength of Corry lies in its rich linguistically motivated feature set, this remains an issue.

	Mention detection			CEAF			MUC			B^3			BLANC		
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1
Language: en, Information: open, Annotation: gold															
Corry-B	100	100	100	77.5	77.5	77.5	56.1	57.5	56.8	82.6	85.7	84.1	69.3	75.3	**71.8**
Corry-C	100	100	100	77.7	77.7	**77.7**	57.4	58.3	57.9	83.1	84.7	83.9	71.3	71.6	71.5
Corry-M	100	100	100	73.8	73.8	73.8	62.5	56.2	**59.2**	85.5	78.6	81.9	76.2	58.8	62.7
RelaxCor	100	100	100	75.8	75.8	75.8	22.6	70.5	34.2	75.2	96.7	**84.6**	58.0	83.8	62.7
Language: en, Information: open, Annotation: regular															
BART	76.1	69.8	72.8	70.1	64.3	67.1	62.8	52.4	57.1	74.9	67.7	71.1	55.3	73.2	57.7
Corry-B	79.8	76.4	**78.1**	70.4	67.4	68.9	55.0	54.2	54.6	73.7	74.1	**73.9**	57.1	75.7	**60.6**
Corry-C	79.8	76.4	**78.1**	70.9	67.9	**69.4**	54.7	55.5	55.1	73.8	73.1	73.5	57.4	63.8	59.4
Corry-M	79.8	76.4	**78.1**	66.3	63.5	64.8	61.5	53.4	**57.2**	76.8	66.5	71.3	58.5	56.2	57.1

Table 1: System scores for the gold/regular open setting. The best F-score for each metric shown in bold.

An important advantage of Corry is its flexibility: the system allows for a number of modeling solutions that can be tested on the development set to optimize the performance level for a particular objective. Our SemEval task 1 results confirm that a system might benefit a lot from a direct optimization for a given performance metric.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at the International Conference on Language Resources and Evaluation (LREC-1998)*, pages 563–566.

Pascal Denis and Jason Baldridge. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics - Human Language Technology Conference (NAACL/HLT-2007)*.

Pascal Denis and Jason Baldridge. 2008. Coreference with named entity classification and transitivity constraints and evaluation with MUC, B-CUBED, and CEAF. In *Proceedings of Corpus-Based Approaches to Coreference Resolution in Romance Languages (CBA 2008)*.

Jenny Rose Finkel and Christopher D. Manning. 2008. Enforcing transitivity in coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008), Short Papers*, pages 45–48.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.

Dan Klein and Christopher Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics - Human Language Technology Conference (NAACL/HLT-2005)*, pages 25–32.

Marta Recasens and Eduard Hovy. in prep. BLANC: Implementing the rand index for coreference 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 Evaluations (SemEval-2010)*, Uppsala, Sweden.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics (Special Issue on Computational Anaphora Resolution)*, 27(4):521–544.

Olga Uryupina. 2003. High-precision identification of discourse-new and unique noun phrases. In *Proceedings of the ACL'03 Student Workshop*, pages 80–86.

Olga Uryupina. 2006. Coreference resolution with and without linguistic knowledge. In *Proceedings of the Language Resources and Evaluation Conference*.

Olga Uryupina. 2007. *Knowledge Acquisition for Coreference Resolution*. Ph.D. thesis, Saarland University.

Olga Uryupina. 2008. Error analysis for learning-based coreference resolution. In *Proceedings of the Language Resources and Evaluation Conference*.

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