# Multigraph Clustering for Unsupervised Coreference Resolution

**Sebastian Martschat**
Heidelberg Institute for Theoretical Studies gGmbH
Schloss-Wolfsbrunnenweg 35
69118 Heidelberg, Germany
sebastian.martschat@h-its.org

## Abstract

We present an unsupervised model for coreference resolution that casts the problem as a clustering task in a directed labeled weighted multigraph. The model outperforms most systems participating in the English track of the CoNLL'12 shared task.

## 1 Introduction

Coreference resolution is the task of determining which mentions in a text refer to the same entity. With the advent of machine learning and the availability of annotated corpora in the mid 1990s the research focus shifted from rule-based approaches to supervised machine learning techniques. Quite recently, however, rule-based approaches regained popularity due to Stanford's multi-pass sieve approach which exhibits state-of-the-art performance on many standard coreference data sets (Raghunathan et al., 2010) and also won the CoNLL-2011 shared task on coreference resolution (Lee et al., 2011; Pradhan et al., 2011). These results show that carefully crafted rule-based systems which employ suitable inference schemes can achieve competitive performance. Such a system can be considered unsupervised in the sense that it does not employ training data for optimizing parameters.

In this paper we present a graph-based approach for coreference resolution that models a document to be processed as a graph. The nodes are mentions and the edges correspond to relations between mentions. Coreference resolution is performed via graph clustering. Our approach belongs to a class of recently proposed graph models for coreference resolution (Cai and Strube, 2010; Sapena et al., 2010; Martschat et al., 2012) and is designed to be a simplified version of existing approaches. In contrast to previous models belonging to this class we do not learn any edge weights but perform inference on the graph structure only which renders our model unsupervised. On the English data of the CoNLL'12 shared task the model outperforms most systems which participated in the shared task.

## 2 Related Work

**Graph-based coreference resolution.** While not developed within a graph-based framework, factor-based approaches for pronoun resolution (Mitkov, 1998) can be regarded as greedy clustering in a multigraph, where edges representing factors for pronoun resolution have negative or positive weight. This yields a model similar to the one presented in this paper though Mitkov's work has only been applied to pronoun resolution. Nicolae and Nicolae (2006) phrase coreference resolution as a graph clustering problem: they first perform pairwise classification and then construct a graph using the derived confidence values as edge weights. In contrast, work by Culotta et al. (2007), Cai and Strube (2010) and Sapena et al. (2010) omits the classification step entirely. Sapena et al. (2010) and Cai and Strube (2010) perform coreference resolution in one step using graph partitioning approaches. These approaches participated in the recent CoNLL'11 shared task (Pradhan et al., 2011; Sapena et al., 2011; Cai et al., 2011b) with excellent results. The approach by Cai et al. (2011b) has been modified by Martschat et al. (2012) and ranked second in the English track at the CoNLL'12 shared task (Pradhan et al., 2012). The top performing system at the CoNLL'12 shared task (Fernandes et al., 2012)



*Proceedings of the ACL Student Research Workshop*, pages 81–88,
Sofia, Bulgaria, August 4-9 2013. 

also represents the problem as a graph by performing inference on trees constructed using the multi-pass sieve approach by Raghunathan et al. (2010) and Lee et al. (2011), which in turn won the CoNLL'11 shared task.

**Unsupervised coreference resolution.** Cardie and Wagstaff (1999) present an early approach to unsupervised coreference resolution based on a straightforward clustering approach. Angheluta et al. (2004) build on their approach and devise more sophisticated clustering algorithms. Haghighi and Klein (2007), Ng (2008) and Charniak and Elsner (2009) employ unsupervised generative models. Poon and Domingos (2008) present a Markov Logic Network approach to unsupervised coreference resolution. These approaches reach competitive performance on gold mentions but not on system mentions (Ng, 2008). The multi-pass sieve approach by Raghunathan et al. (2010) can also be viewed as unsupervised.

## 3 A Multigraph Model

We aim for a model which directly represents the relations between mentions in a graph structure. Clusters in the graph then correspond to entities.

### 3.1 Motivation

To motivate the choice of our model, let us consider a simple made-up example.

> *Leaders* met in *Paris* to discuss *recent developments*. *They* left the city today.

We want to model that *Paris* is not a likely candidate antecedent for *They* due to number disagreement, but that *Leaders* and *recent developments* are potential antecedents for *They*. We want to express that *Leaders* is the preferred antecedent, since *Leaders* and *They* are in a parallel construction both occupying the subject position in their respective sentences.

In other words, our model should express the following relations for this example:

- number disagreement for (*They*, *Paris*), which indicates that the mentions are not coreferent,
- the anaphor being a pronoun for (*They*, *Leaders*), (*They*, *recent developments*) and (*They*, *Paris*), which is a weak indicator for coreference if the mentions are close to each other,
- syntactic parallelism for (*They*, *Leaders*): both mentions are in a parallel construction in adjacent sentences (both in the subject slot), which is also a weak coreference indicator.

We denote these relations as N_Number, P_AnaPron and P_Subject respectively. The graphical structure depicted in Figure 1 models these relations between the four mentions *Leaders*, *Paris*, *recent developments* and *They*.

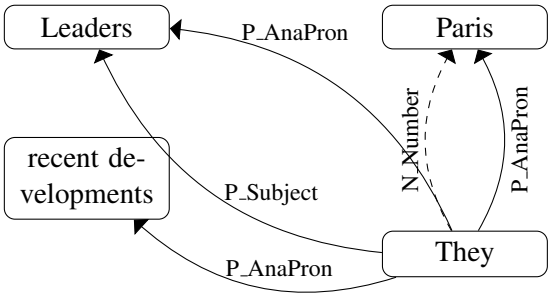


Figure 1: An example graph modeling relations between mentions.

A directed edge from a mention $m$ to $n$ indicates that $n$ precedes $m$ and that there is some relation between $m$ and $n$ that indicates coreference or non-coreference. Labeled edges describe the relations between the mentions, multiple relations can hold between a pair. Edges may be weighted.

### 3.2 Multigraphs for Coreference Resolution

Formally, the model is a *directed labeled weighted multigraph*. That is a tuple $D = (R, V, A, w)$ where

- $R$ is the set of labels (in our case relations such as P_Subject that hold between mentions),
- $V$ is the set of nodes (the mentions extracted from a document),
- $A \subseteq V \times V \times R$ is the set of edges (relations between two mentions),
- $w$ is a mapping $w\colon\ A \to \mathbb{R} \cup \{\pm\infty\}$ (weights for edges).

Many graph models for coreference resolution operate on $A = V \times V$. Our multigraph model allows us to have multiple edges with different labels between mentions.

To have a notion of *order* we employ a directed graph: We only allow an edge from $m$ to $n$ if $m$ appears later in the text than $n$.

To perform coreference resolution for a document $d$, we first construct a directed labeled multigraph (Section 3.3). We then assign a weight to each edge (Section 3.4). The resulting graph is

clustered to obtain the mentions that refer to the same entity (Section 3.5).

### 3.3 Graph Construction

Given a set $M$ of mentions extracted from a document $d$, we set $V = M$, i.e. the nodes of the graph are the mentions. To construct the edges $A$, we consider each pair $(m, n)$ of mentions with $n \prec m$. We then check for every relation $r \in R$ if $r$ holds for the pair $(m, n)$. If this is the case we add the edge $(m, n, r)$ to $A$. For simplicity, we restrict ourselves to binary relations that hold between pairs of mentions (see Section 4).

The graph displayed in Figure 1 is the graph constructed for the mentions *Leaders*, *Paris*, *recent developments* and *They* from the example sentence at the beginning of this Section, where $R = \{$P_AnaPron, P_Subject, N_Number$\}$.

### 3.4 Assigning Weights

Depending on whether a relation $r \in R$ is indicative for non-coreference (e.g. number disagreement) or for coreference (e.g. string matching) it should be weighted differently. We therefore divide $R$ into a set of *negative relations* $R_-$ and a set of *positive relations* $R_+$.

Previous work on multigraphs for coreference resolution disallows any edge between mentions for which a negative relations holds (Cai et al., 2011b; Martschat et al., 2012). We take a similar approach and set $w(m, n, r) = -\infty$ for $(m, n, r) \in A$ when $r \in R_-$[1].

Work on graph-based models similar to ours report robustness with regard to the amount of training data used (Cai et al., 2011b; Cai et al., 2011a; Martschat et al., 2012). Motivated by their observations we treat every positive relation equally and set $w(m, n, r) = 1$ for $(m, n, r) \in A$ if $r \in R_+$.

In contrast to previous work on similar graph models we do not learn any edge weights from training data. We compare this unsupervised scheme with supervised variants empirically in Section 5.

### 3.5 Clustering

To describe the clustering algorithm used in this work we need some additional terminology. If there exists an edge $(m, n, r) \in A$ we say that $n$ is a *child* of $m$.

[1]We experimented with different weighting schemes for negative relations on development data (e.g. setting $w(m, n, r) = -1$) but did not observe a gain in performance.

In the graph constructed according to the procedure described in Section 3.3, all children of a mention $m$ are candidate antecedents for $m$. The relations we employ are indicators for coreference (which get a positive weight) and indicators for non-coreference (which get a negative weight). We aim to employ a simple and efficient clustering scheme on this graph and therefore choose 1-nearest-neighbor clustering: for every $m$, we choose as antecedent $m$'s child $n$ such that the sum of edge weights is maximal and positive. We break ties by choosing the closest mention.

In the unsupervised setting described in Section 3.4 this algorithm reduces to choosing the child that is connected via the highest number of positive relations and via no negative relation.

For the graph depicted in Figure 1 this algorithm computes the clusters $\{$*They*, *Leaders*$\}$, $\{$*Paris*$\}$ and $\{$*recent developments*$\}$.

## 4 Relations

The graph model described in Section 3 is based on expressing relations between pairs of mentions via edges built from such relations. We now describe the relations currently used by our system. They are well-known indicators and constraints for coreference and are taken from previous work (Cardie and Wagstaff, 1999; Soon et al., 2001; Rahman and Ng, 2009; Lee et al., 2011; Cai et al., 2011b). All relations operate on pairs of mentions $(m, n)$, where $m$ is the anaphor and $n$ is a candidate antecedent. If a relation $r$ holds for $(m, n)$, the edge $(m, n, r)$ is added to the graph. We finalized the set of relations and their distance thresholds on development data.

### 4.1 Negative Relations

Negative relations receive negative weights. They allow us to introduce well-known constraints such as agreement into our model.

**(1) N_Gender, (2) N_Number:** Two mentions do not agree in gender or number. We compute number and gender for common nouns using the number and gender data provided by Bergsma and Lin (2006).

**(3) N_SemanticClass:** Two mentions do not agree in semantic class (we only use the top categories *Object*, *Date* and *Person* from WordNet (Fellbaum, 1998)).

**(4) N_ItDist:** The anaphor is *it* or *they* and the sentence distance to the antecedent is larger

than one.

**(5) N_Speaker12Pron:** Two first person pronouns or two second person pronouns with different speakers, or one first person pronoun and one second person pronoun with the same speaker[2].

**(6) N_ContraSubObj:** Two mentions are in the subject/object positions of the same verb, the anaphor is a non-possessive/reflexive pronoun.

**(7) N_Mod:** Two mentions have the same syntactic heads, and the anaphor has a nominal modifier which does not occur in the antecedent.

**(8) N_Embedding:** Two mentions where one embeds the other, which is not a reflexive or possessive pronoun.

**(9) N_2PronNonSpeech:** Two second person pronouns without speaker information and not in direct speech.

### 4.2 Positive Relations

Positive relations are coreference indicators which are added as edges with positive weights.

**(10) P_NonPron_StrMatch:** Applies only if the anaphor is definite or a proper name[3]. This relation holds if after discarding stop words the strings of mentions completely match.

**(11) P_HeadMatch:** If the syntactic heads of mentions match.

**(12) P_Alias:** If mentions are aliases of each other (i.e. proper names with partial match, full names and acronyms, etc.).

**(13) P_Speaker12Pron:** If the speaker of the second person pronoun is talking to the speaker of the first person pronoun (applies only to first/second person pronouns).

**(14) P_DSPron:** One mention is a *speak verb*'s subject, the other mention is a first person pronoun within the corresponding direct speech.

**(15) P_ReflPronSub:** If the anaphor is a reflexive pronoun, and the antecedent is the subject of the sentence.

**(16) P_PossPronSub:** If the anaphor is a possessive pronoun, and the antecedent is the subject of the anaphor's sentence or subclause.

**(17) P_PossPronEmb:** The anaphor is a possessive pronoun embedded in the antecedent.

**(18) P_AnaPron:** If the anaphor is a pronoun and none of the mentions is a first or second person pronoun. This relation is restricted to a sentence distance of 3.

**(19) P_VerbAgree:** If the anaphor is a third person pronoun and has the same predicate as the antecedent. This relation is restricted to a sentence distance of 1.

**(20) P_Subject, (21) P_Object:** The anaphor is a third person pronoun and both mentions are subjects/objects. These relations are restricted to a sentence distance of 1.

**(22) P_Pron_StrMatch:** If both mentions are pronouns and their strings match.

**(23) P_Pron_Agreement:** If both mentions are different pronoun tokens but agree in number, gender and person.

## 5 Evaluation

### 5.1 Data and Evaluation Metrics

We use the data provided for the English track of the CoNLL'12 shared task on multilingual coreference resolution (Pradhan et al., 2012) which is a subset of the upcoming OntoNotes 5.0 release and comes with various annotation layers provided by state-of-the-art NLP tools. We used the official dev/test split for development and evaluation. We evaluate the model in a setting that corresponds to the shared task's *closed track*, i.e. we use only WordNet (Fellbaum, 1998), the number and gender data of Bergsma and Lin (2006) and the provided annotation layers. To extract system mentions we employ the mention extractor described in Martschat et al. (2012).

We evaluate our system with the coreference resolution evaluation metrics that were used for the CoNLL shared tasks on coreference, which are MUC (Vilain et al., 1995), $B^3$ (Bagga and Baldwin, 1998) and $CEAF_e$ (Luo, 2005). We also report the unweighted *average* of the three scores, which was the official evaluation metric in the shared tasks. To compute the scores we employed the official scorer supplied by the shared task organizers.

### 5.2 Results

Table 1 displays the performance of our model and of the systems that obtained the best (Fernandes et al., 2012) and the median performance in the

[2]Like all relations using speaker information, this relation depends on the gold speaker annotation layer in the corpus.

[3]This condition is necessary to cope with the high-recall output of the mention tagger.

| | MUC | | | $B^3$ | | | $CEAF_e$ | | | average |
|---|---|---|---|---|---|---|---|---|---|---|
| | R | P | F1 | R | P | F1 | R | P | F1 | |
| **CoNLL'12 English development data** | | | | | | | | | | |
| best | 64.88 | 74.74 | 69.46 | 66.53 | 78.28 | 71.93 | 54.93 | 43.68 | 48.66 | 63.35 |
| median | 62.3 | 62.8 | 62.0 | 66.7 | 71.8 | 69.1 | 46.4 | 44.9 | 45.6 | 58.9 |
| this work (weights_fraction) | 64.00 | 68.56 | 66.20 | 66.59 | 75.67 | 70.84 | 50.48 | 45.52 | 47.87 | 61.63 |
| this work (weights_MaxEnt) | 63.72 | 65.78 | 64.73 | 66.60 | 73.76 | 70.00 | 47.46 | 45.30 | 46.36 | 60.36 |
| this work (unsupervised) | 64.01 | 68.58 | 66.22 | 67.00 | 76.45 | 71.41 | 51.10 | 46.16 | 48.51 | 62.05 |
| **CoNLL'12 English test data** | | | | | | | | | | |
| best | 65.83 | 75.91 | 70.51 | 65.79 | 77.69 | 71.24 | 55.00 | 43.17 | 48.37 | 63.37 |
| median | 62.08 | 63.02 | 62.55 | 66.23 | 70.45 | 68.27 | 45.74 | 44.74 | 45.23 | 58.68 |
| this work (weights_fraction) | 64.25 | 68.31 | 66.22 | 65.44 | 74.20 | 69.54 | 49.18 | 44.71 | 46.84 | 60.87 |
| this work (weights_MaxEnt) | 63.58 | 64.70 | 64.14 | 65.63 | 72.09 | 68.71 | 45.58 | 44.41 | 44.99 | 59.28 |
| this work (unsupervised) | 63.95 | 67.99 | 65.91 | 65.47 | 74.93 | 69.88 | 49.83 | 45.40 | 47.51 | 61.10 |

Table 1: Results of different systems on the CoNLL'12 English data sets.

CoNLL'12 shared task, which are denoted as *best* and *median* respectively. *best* employs a structured prediction model with learned combinations of 70 basic features. We also compare with two supervised variants of our model which use the same relations and the same clustering algorithm as the unsupervised model: *weights_fraction* sets the weight of a relation to the fraction of positive instances in training data (as in Martschat et al. (2012)). *weights_MaxEnt* trains a mention-pair model (Soon et al., 2001) via the maximum entropy classifier implemented in the BART toolkit (Versley et al., 2008) and builds a graph where the weight of an edge connecting two mentions is the classifier's prediction[4]. We use the official CoNLL'12 English training set for training.

Our unsupervised model performs considerably better than the median system from the CoNLL'12 shared task on both data sets according to all metrics. It also seems to be able to accommodate well for the relations described in Section 4 since it outperforms both supervised variants[5]. The model performs worse than *best*, the gap according to $B^3$ and $CEAF_e$ being considerably smaller than according to MUC. While we observe a decrease of 1 point average score when evaluating on test data the model still would have ranked fourth in the English track of the CoNLL'12 shared task with only 0.2 points difference in average score to the second ranked system.

[4]The classifier's output is a number $p \in [0, 1]$. In order to have negative weights we use the transformation $p' = 2p - 1$.

[5]Compared with the supervised variants all improvements in F1 score are statistically significant according to a paired t-test ($p < 0.05$) except for the difference in MUC F1 to *weights_fraction*.

## 6 Error Analysis

In order to understand weaknesses of our model we perform an error analysis on the development data. We distinguish between *precision* and *recall* errors. For an initial analysis we split the errors according to the mention type of anaphor and antecedent (name, nominal and pronoun).

### 6.1 Precision Errors

Our system operates in a pairwise fashion. We therefore count one precision error whenever the clustering algorithm assigns two non-coreferent mentions to the same cluster. Table 2 shows the number of clustering decisions made according to the mention type and in brackets the fraction of decisions that erroneously assign two non-coreferent mentions to the same cluster. We see that two main sources of error are nominal-nominal pairs and the resolution of pronouns. We now focus on gaining further insight into the system's performance for pronoun resolution by investigating the performance per pronoun type. The results are displayed in Table 3. We obtain good performance for *I* and *my* which in the majority of cases can be resolved unambiguously by the speaker relations employed by our system. The relations we use also seem

| | NAM | NOM | PRO |
|---|---|---|---|
| NAM | 3413 (21%) | 67 (66%) | 11 (46%) |
| NOM | 43 (67%) | 2148 (49%) | 9 (89%) |
| PRO | 868 (32%) | 1771 (55%) | 5308 (24%) |

Table 2: Number of clustering decisions made according to mention type (rows anaphor, columns antecedent) and percentage of wrong decisions.

| Anaphor | all | anaphoric |
|---|---|---|
| I | 1260 (13%) | 1239 (11%) |
| my | 192 (14%) | 181 (9%) |
| he | 824 (14%) | 812 (13%) |
| ... | ... | |
| they | 764 (29%) | 725 (26%) |
| ... | ... | |
| you | 802 (41%) | 555 (15%) |
| it | 1114 (64%) | 720 (44%) |

Table 3: Precision statistics for pronouns. Rows are pronoun surfaces, columns number of clustering decisions and percentage of wrong decisions for all and only anaphoric pronouns respectively.

to work well for *he*. In contrast, the local, shallow approach we currently employ is not able to resolve highly ambiguous pronouns such as *they*, *you* or *it* in many cases. The reduction in error rate when only considering anaphoric pronouns shows that our system could benefit from an improved detection of expletive *it* and *you*.

### 6.2 Recall Errors

Estimating recall errors by counting all missing pairwise links would consider each entity many times. Therefore, we instead count one recall error for a pair $(m, n)$ of anaphor $m$ and antecedent $n$ if (i) $m$ and $n$ are coreferent, (ii) $m$ and $n$ are not assigned to the same cluster, (iii) $m$ is the first mention in its cluster that is coreferent with $n$, and (iv) $n$ is the closest mention coreferent with $m$ that is not in $m$'s cluster.

This can be illustrated by an example. Considering mentions $m_1, \ldots, m_5$, assume that $m_1$, $m_3$, $m_4$ and $m_5$ are coreferent but the system clusters are $\{m_2, m_3\}$ and $\{m_4, m_5\}$. We then count two recall errors: one for the missing link from $m_3$ to $m_1$ and one for the missing link from $m_4$ to $m_3$.

According to this definition we count 3528 recall errors on the development set. The distribution of errors is displayed in Table 4. We see that

| | NAM | NOM | PRO |
|---|---|---|---|
| NAM | 321 | 220 | 247 |
| NOM | 306 | 797 | 330 |
| PRO | 306 | 476 | 525 |

Table 4: Number of recall errors according to mention type (rows anaphor, columns antecedent).

the main source of recall errors are missing links of nominal-nominal pairs. We randomly extracted 50 of these errors and manually assigned them to different categories.

**29 errors: missing semantic knowledge**. In these cases lexical or world knowledge is needed to build coreference links between mentions with different heads. For example our system misses the link between *the sauna* and *the hotbox sweatbox*.

**14 errors: too restrictive N_Mod**. In these cases the heads of the mentions matched but no link was built due to N_Mod. An example is the missing link between *our island's last remaining forest of these giant trees* and *the forest of Chilan*.

**4 errors: too cautious string match**. We only apply string matching for common nouns when the noun is definite.

Three errors could not be attributed to any of the above categories.

## 7 Conclusions and Future Work

We presented an unsupervised graph-based model for coreference resolution. Experiments show that our model exhibits competitive performance on the English CoNLL'12 shared task data sets.

An error analysis revealed that two main sources of errors of our model are the inaccurate resolution of highly ambiguous pronouns such as *it* and missing links between nominals with different heads. Future work should investigate how semantic knowledge and more complex relations capturing deeper discourse properties such as coherence or information status can be added to the model. Processing these features efficently may require a more sophisticated clustering algorithm.

We are surprised by the good performance of this unsupervised model in comparison to the state-of-the-art which uses sophisticated machine learning techniques (Fernandes et al., 2012) or well-engineered rules (Lee et al., 2011). We are not sure how to interpret these results and want to leave different interpretations for discussion:

- our unsupervised model is really that good (hopefully),
- the evaluation metrics employed are to be questioned (certainly),
- efficiently making use of annotated training data still remains a challenge for the state-of-the-art (likely).

## Acknowledgments

This work has been funded by the Klaus Tschira Foundation, Germany. The author has been supported by a HITS PhD scholarship.

## References


Roxana Angheluta, Patrick Jeuniaux, Rudradeb Mitra, and Marie-Francine Moens. 2004. Clustering algorithms for noun phrase coreference resolution. In *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles,* Louvain La Neuve, Belgium, 10–12 March 2004, pages 60–70.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the 1st International Conference on Language Resources and Evaluation,* Granada, Spain, 28–30 May 1998, pages 563–566.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics,* Sydney, Australia, 17–21 July 2006, pages 33–40.

Jie Cai and Michael Strube. 2010. End-to-end coreference resolution via hypergraph partitioning. In *Proceedings of the 23rd International Conference on Computational Linguistics,* Beijing, China, 23–27 August 2010, pages 143–151.

Jie Cai, Éva Mújdricza-Maydt, Yufang Hou, and Michael Strube. 2011a. Weakly supervised graph-based coreference resolution for clinical data. In *Proceedings of the 5th i2b2 Shared Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data,* Washington, D.C., 20-21 October 2011.

Jie Cai, Éva Mújdricza-Maydt, and Michael Strube. 2011b. Unrestricted coreference resolution via global hypergraph partitioning. In *Proceedings of the Shared Task of the 15th Conference on Computational Natural Language Learning,* Portland, Oreg., 23–24 June 2011, pages 56–60.

Claire Cardie and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. In *Proceedings of the 1999 SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora,* College Park, Md., 21–22 June 1999, pages 82–89.

Eugene Charniak and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics,* Athens, Greece, 30 March – 3 April 2009, pages 148–156.

Aron Culotta, Michael Wick, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics,* Rochester, N.Y., 22–27 April 2007, pages 81–88.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Eraldo Fernandes, Cícero dos Santos, and Ruy Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Proceedings of the Shared Task of the 16th Conference on Computational Natural Language Learning,* Jeju Island, Korea, 12–14 July 2012, pages 41–48.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,* Prague, Czech Republic, 23–30 June 2007, pages 848–855.

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 Shared Task of the 15th Conference on Computational Natural Language Learning,* Portland, Oreg., 23–24 June 2011, pages 28–34.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing,* Vancouver, B.C., Canada, 6–8 October 2005, pages 25–32.

Sebastian Martschat, Jie Cai, Samuel Broscheit, Éva Mújdricza-Maydt, and Michael Strube. 2012. A multigraph model for coreference resolution. In *Proceedings of the Shared Task of the 16th Conference on Computational Natural Language Learning,* Jeju Island, Korea, 12–14 July 2012, pages 100–106.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics,* Montréal, Québec, Canada, 10–14 August 1998, pages 869–875.

Vincent Ng. 2008. Unsupervised models for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, Honolulu, Hawaii, 25–27 October 2008, pages 640–649.

Cristina Nicolae and Gabriel Nicolae. 2006. BestCut: A graph algorithm for coreference resolution. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,* Sydney, Australia, 22–23 July 2006, pages 275–283.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, Honolulu, Hawaii, 25–27 October 2008, pages 650–659.

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 Shared Task of the 15th Conference on Computational Natural Language Learning,* Portland, Oreg., 23–24 June 2011, pages 1–27.

Sameer Pradhan, Alessandro Moschitti, and Nianwen Xue. 2012. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Shared Task of the 16th Conference on Computational Natural Language Learning,* Jeju Island, Korea, 12–14 July 2012, pages 1–40.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing,* Cambridge, Mass., 9–11 October 2010, pages 492–501.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* Singapore, 6–7 August 2009, pages 968–977.

Emili Sapena, Lluís Padró, and Jordi Turmo. 2010. A global relaxation labeling approach to coreference resolution. In *Proceedings of Coling 2010: Poster Volume,* Beijing, China, 23–27 August 2010, pages 1086–1094.

Emili Sapena, Lluís Padró, and Jordi Turmo. 2011. RelaxCor participation in CoNLL shared task on coreference resolution. In *Proceedings of the Shared Task of the 15th Conference on Computational Natural Language Learning,* Portland, Oreg., 23–24 June 2011, pages 35–39.

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(4):521–544.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: A modular toolkit for coreference resolution. In *Companion Volume to the Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics,* Columbus, Ohio, 15–20 June 2008, pages 9–12.

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 (MUC-6)*, pages 45–52, San Mateo, Cal. Morgan Kaufmann.