

# A SURVEY OF MACHINE LEARNING FOR REFERENCE RESOLUTION IN TEXTUAL DISCOURSE

Fredrik Olsson  
fredrik.olsson@sics.se

December 18, 2003

## 1. Introduction

Machine learning methods have been successfully applied to a number of language technology tasks. Previous work has to a large extent been focused on intra-sentential phenomenon, such as part of speech tagging, phrase chunking and named entity recognition. The purpose of this report is to investigate the efforts made in applying machine learning methods for handling the *inter*-sentential linguistic phenomenon of anaphoric reference and coreference in text. Due to the amount of literature available on the subject, this report will not present a paper-by-paper survey, but rather try to synthesise the information available in the sources and line up general observations. When necessary, individual approaches will be examined and disseminated. On a general note, Mitkov (2002) provides a point of departure for the computational treatment of anaphora resolution.

The term *reference resolution* in the title of this report is deliberately somewhat vague as it covers both coreference and anaphora<sup>1</sup>. While coreference is defined as a relation holding between noun phrases if they refer to the same entity (see e.g. (Hirschman and Chinchor, 1997)), anaphora is understood as the presupposition of something that has gone before and that points back to some previous item (Halliday and Hasan, 1976). Thus, there is a difference between anaphora and coreference, and, as van Deemter and Kibble (2000) point out; coreference is an equivalence relation, while anaphora is irreflexive, non-symmetrical and non-transitive relation. This imply that the phenomenon of anaphora is sensitive to context, while coreference is not.

A number of approaches has been proposed for English noun phrase reference resolution, e.g., McCarthy and Lehnert (1995), Cardie and Wagstaff (1999), Soon *et al* (2001), Harabagiu *et al* (2001), Preiss (2002), Ng and Cardie (2002c; 2002b; 2002a; 2003a; 2003b), Yang *et al* (2003), as well as for Japanese noun phrases, e.g. Lida *et al* (2003), and German, e.g. Strube *et al* (2002), Müller *et al* (2002). Some attempts at solving English

anaphora is reported by, e.g. Connolly *et al* (1997), and Ge *et al* (1998), while Evans (2001) report on experiments for automatic classification of *it*. Aone and Bennett (1995) report on work on Japanese, especially the treatment of zero-pronouns. Modejska *et al* (2003) report on experiments done on solving other-anaphora, that is, referential noun phrases modified by “other” or “another” and having non-structural antecedents. Finally, Soderland and Lehnert (1994a; 1994b) describe the implementation of a system for inter-sentential reference generation in the setting of an information extraction system.

## 2. Recasting the problem

The problem of resolving references is often perceived as identifying chains of referents, often across sentence boundaries, and sometimes even across documents. To make the problem more manageable for machine learning algorithms, it is often recasted as a classification and a clustering task. A classifier determines whether or not two potential referents are coreferent or anaphoric, and a clustering mechanism coordinates the pairwise coreferent items into partitions, each of which contains the items that refer to the same entity.

Since coreference is an equivalence relation, and anaphora is not, the two types of reference need somewhat different treatment. Coarsely, it can be argued that clustering is always performed implicitly in the case of anaphoric reference resolution, that is, since one is bound to look for an antecedent noun phrase rather than a pronoun that points to a noun phrase, what is obtained is a cluster of pronouns referring to the same noun phrase, rather than a chain of referents pointing back to a noun phrase. Thus anaphora resolution, as reported on by Aone and Bennett (1995), Connolly *et al* (1997), and Ge *et al* (1998), may be seen as a special case of coreference. In the following sections, coreference is what will be focused on.

### 2.1. Classification

Recasting the reference resolution problem as a classification task introduces a couple of new problems. First, it is necessary to know *what* in the input text is to be considered as candidates for classification. Second, given that we know about what kind of entities we

---

<sup>1</sup>The related concept of cataphora is rarely dealt with; in the present survey, it was found briefly mentioned by Stuckardt (2002) and by Evans (2001).

have at our disposal, *how* do we choose an appropriate set to feed to the classifier?

Textual entities participating in a coreference relation are known as *markables* (Soon et al., 2001). Markables, which constitutes the candidates for classification, can be, e.g., definite noun phrases, demonstrative noun phrases, or proper names. In theory, a precondition for classification is to obtain all markables and nothing but the markables present in the input text, i.e., all noun phrases participating in coreference relations in the document. In practice, however, the input to the reference resolver is often far from perfect; the identification of markables rely on non-perfect up-stream components such as part-of-speech taggers and phrase chunkers, each introducing an error which might have grown large by the time the input is prepared for the reference resolver (Mitkov, 2001).

To assess that a machine learning algorithm under investigation constitutes a plausible route toward a solution, it is sometimes necessary to ignore the error-chaining effect and assume that the input to the classifier is perfect. This is done by, e.g., Soderland and Lehnert (1994a; 1994b), who describe an approach involving interacting decision trees (the ID3 algorithm; see, e.g., (Mitchell, 1997)), aiming at inter-sentential inference generation implemented in a system called Wrap-Up. Soderland and Lehnert (1994a; 1994b) establish that their approach serves its purpose; to increase portability and reduce development time without degradation of overall system performance. In settings like these, where a non-learning module is replaced by a learning one in an existing system (an information extraction system in the case of Wrap-Up), it is possible to assess the contribution made by the learning module w.r.t. the relative performance of the non-learning module. Then, for the sake of relative performance, it is possible to ignore the error-chaining effect and still conduct a meaningful evaluation. Experiments along these lines were carried out by, e.g., McCarthy and Lehnert (1995) who, as Soderland and Lehnert (1994a; 1994b), performed comparative evaluations of non-learning and learning reference resolution systems using the MUC-5 corpora. The learning component performed on par with the non-learning one in these experiments.

The other, perhaps more pragmatic approach to reference resolution, is one where the negative effect of error-chaining is not neglected. For instance, Soon et al (2001) present an experiment in which the performance of the learning reference resolution components is not contrasted with a non-learning component in the same process pipeline. Ng and Cardie (2002a) take a different approach and employ a pre-processing component for deciding whether or not a given noun phrase is anaphoric, before submitting it to the coreference classifier.

Once the markables in the input are known, it is possible to adopt an approach where, for every noun

phrase  $A$  for which we look for antecedents or corefering noun phrases, all combinations of  $A$  and a markable are submitted to the coreference classifier. However, this may lead to undesired increases in time and memory required for classification. A solution would be to pick out a smaller candidate set of markables to combine with  $A$ . Most of the papers in this survey does not concern themselves with this issue. Others do, for instance, Yang et al (2003) elaborate on a candidate filter that “eliminate the invalid or irrelevant candidates”, while Ge et al (1998) make use of Hobb’s algorithm for pronoun resolution (Hobbs, 1986) in order to obtain a set of candidate antecedents.

When the markables have been identified (and an appropriate set of candidates have been established), the scene is set for a classifier to pick out referents. There are several ways of obtaining a classifier for reference resolution, and the kind of classifier suitable depends on, among other things, the training data available (see Section 4.) and how data is represented (see Section 3.). Table 1 outlines the machine learning algorithms used in the surveyed papers.

## 2.2. Clustering

The output from a classifier is often a set of pairs of corefering noun phrases. In order to form coreference chains, these pairs must be grouped together, each group representing a coreference chain. As pointed out by Ng and Cardie (2002a), realising the problem of coreference in terms of pair-wise classification of candidate noun phrases, makes impossible the enforcement of the transitivity constraint inherent in coreference. For example, a classifier might classify noun phrases  $A$  and  $B$  as coreferent, and  $B$  and  $C$  as coreferent, but it might also conclude that  $A$  and  $C$  are not coreferent. This problem is present in the single-link clustering algorithm (described by, e.g., Manning and Schütze (1999)), which is commonly used in the reference resolution. Single-link clustering produce clusters that are locally coherent. However, the clusters obtained may be elongated due to what is called the chaining effect which occurs when a chain of large similarities is exploited without taking into account the global context. As a result, the distance between two objects in the same cluster may be greater than the distance between one of the objects and an object situated outside the cluster.

A couple of alternative clustering algorithms are proposed in the literature. Kehler (1997) utilise Dempster’s Rule of Combination to combine pairwise, possibly contradictory, probability distributions into a third probability distribution that represent the consensus of the original two. This way, Kehler manages to assign probabilities to alternative sets of coreference relations produced by an existing, rule based information extraction system.

Ng and Cardie (2002c; 2002a) propose a best-first clustering algorithm that find the most likely corefer-

Machine learning method	Author(s), implementation (where applicable)
Decision trees (Mitchell, 1997)	Soderland and Lehnert (1994a; 1994b), ID3 McCarthy and Lehnert (1995), C4.5 Aone and Bennett (1995), C4.5 Connolly <i>et al</i> (1997), C4.5 Stuckardt (2002), C4.5 Ng and Cardie (2002c; 2002b), C4.5 Soon <i>et al</i> (2001), C5.0 Yang <i>et al</i> (2003), C5.0 Strube <i>et al</i> (2002), J48 Müller <i>et al</i> (2002), J48
Naïve Bayes (Mitchell, 1997)	Connolly <i>et al</i> (1997) Ge <i>et al</i> (1998) Ng and Cardie (2003a; 2003b) Modejska <i>et al</i> (2003)
Rule learning (Cohen, 1995) and (Harabagiu <i>et al.</i> , 2001)	Ng and Cardie (2002c; 2002b; 2002a), Ripper Harabagiu <i>et al</i> (2001)
Genetic algorithms (Mitchell, 1997)	Byron and Allen (1999) Evans (2002)
Memory-based learning (Mitchell, 1997)	Evans (2001), TiMBL Preiss (2002), TiMBL
Maximum entropy (Berger <i>et al.</i> , 1996)	Kehler (1997)
Neural networks (Mitchell, 1997)	Connolly <i>et al</i> (1997)
Clustering (Manning and Schütze, 1999)	Cardie and Wagstaff (1999)
Decision lists (Collins and Singer, 1999)	Ng and Cardie (2003a)
Support vector machines (Boser <i>et al.</i> , 1992) and (Cortes and Vapnik, 1995)	Lida <i>et al</i> (2003)
Co-training (Blum and Mitchell, 1998)	Müller <i>et al</i> (2002) Ng and Cardie (2003a; 2003b)
Self-training (Banko and Brill, 2001)	Ng and Cardie (2003b)
Expectation-Maximization (Nigam <i>et al.</i> , 2000)	Ng and Cardie (2003b)

Table 1: Machine learning methods used in the surveyed papers. References in the left column indicate sources of information on a particular machine learning method.

ent noun phrase, rather than the first coreferent noun phrase that would be obtained using a single-link clusterer.

### 2.3. Classification as clustering

There are, of course, exceptions to the classification/clustering approach to reference resolution outlined in Section 2. Cardie and Wagstaff (1999) view noun phrase resolution as a pure clustering problem and introduce an unsupervised, greedy clustering algorithm for partitioning noun phrases into equivalence classes. The algorithm starts at the end of the input text, and works toward the beginning. Each noun phrase is compared to all preceding noun phrases, and if the distance, according to some metric, is less than a given threshold, the classes of the noun phrases compared are considered for merging. Two classes can

be merged unless they contain any incompatible noun phrases. Roughly, the incompatibility measure is defined as a function for each feature used in representing the noun phrases.

The performance reported by Cardie and Wagstaff (1999) places their approach in the middle of the field when evaluated on the MUC-6 coreference resolution data set.

### 2.4. Hybrid approaches

There are also efforts reported on combining non-learning coreference resolution systems with learning ones. Kehler (1997) reports on a method for assigning probability distributions to alternative sets of corefering entities in the output of a non-learning information extraction system. Kehler envisions a scenario in which the extraction system is only one of several

sources of information. A downstream system must then be able to combine the, possibly contradictory, evidence delivered from the information sources. To accommodate for this, each source must be assigned the probability that the information it delivers is reliable according to some criterion. This way, the downstream system can disregard unreliable information from one source in favor of more reliable information produced by another.

Byron and Allen (1999) present a very short report on using genetic algorithms for automatically assigning weights to the factors contributing to a non-learning pronoun resolution system’s overall performance.

Genetic algorithms are also employed by Evans (2002), who present work on modifying an existing non-learning pronominal anaphora resolution system to account for different types of pronouns being treated in different ways. The working hypothesis is that the performance of the non-learning system will improve if the system is allowed to behave differently when resolving different types of pronouns. Evans (2002) report that the approach improved the performance of the original non-learning system for some particular texts but not all for all texts in general. Evans points out that there was not enough representative data available for the genetic algorithms to train properly, but on the other hand he states “Unfortunately, the more data is available, the less likely it is that the GA will find values to obtain increased performance” (Evans, 2002).

Finally, Stuckardt (2002) combines a robust general-domain non-learning anaphora resolution system with a learning system that contributes domain knowledge. The combined system is found to perform on par with the original non-learning system.

### 3. Features

Features are the means by which the characteristics of the data to learn from are described to a machine learning algorithm. As noted by, e.g., Stuckardt (2002), finding an appropriate set of features that is hard and time consuming. Not only should the features reflect the concept under investigation in an elaborate (and sometime even redundant) manner, the features should also be computationally inexpensive to obtain from the input data. Tables 2 and 3, contain an outline of the features and data used for anaphoric, and coreference resolution, respectively, covered by the papers under investigation.

For anaphora resolution, the number features reported on in the surveyed papers range from 4 (Ge et al., 1998) to 66 (Aone and Bennett, 1995). The same figures for coreference resolution are 8 (McCarthy and Lehnert, 1995) and 53 (Ng and Cardie, 2002c).<sup>2</sup> The

---

<sup>2</sup>The attentive reader have noticed that Kehler (1997) exploits only three characteristics. However, the task Kehler investigates is assigning probability distributions to

type of features used are often related to the immediate local context of the markables to classify, and sometime to the pair of markables under consideration. Walker (1989), who manually compared two linguistically motivated algorithms for anaphora resolution ((Hobbs, 1986), (Brennan et al., 1987)), found that global focus, cue words and syntax are important factors to discourse components. In general, the features used are not grounded in linguistic theories about reference. One exception is reported by Lida *et al* (2003) who show that centering features contribute to better performance of their coreference resolution system for Japanese.

## 4. Data

Obtaining training and test data is generally a bottleneck to most supervised machine learning approaches, and it seems to constitute a major obstacle to learning systems dealing with high level abstraction of linguistic phenomenon in particular. One problem is the amount of annotated text needed in order to properly cover high level abstractions such as reference resolution; the higher the abstraction level, the higher the risk of data sparseness. An annotator simply has to process much more text to cover enough instances of reference resolution than needed for, say, part of speech tagging. Another problem is the fact that humans do not always agree on how to define and mark high level linguistic phenomenon in text. For instance, when humans annotate texts for coreference to be used for training and testing, the inter-annotator agreement has been found to be rather low; in a study conducted on the MUC-6 and MUC-7 data sets used in evaluating information extraction systems, Hirschman *et al* (1997) found the inter-annotator agreement to be in the lower 90’s for the task of annotating coreference.

### 4.1. Available corpora

Several of the authors of the papers underlying this report, acknowledge the need for more publicly available data annotated for anaphora and coreference. As can be seen in tables 2 and 3, the MUC-6 and MUC-7 corpora are popular. They have now grown to become a de-facto gold standard by which researchers are able to compare their results. The MUC-6 and MUC-7 corpora are both available for purchase from the Linguistic Data Consortium (Chinchor and Sundheim, 1996; Chinchor and Sundheim, 2003), (Linguistic Data Consortium, 2001).

### 4.2. From annotated corpora to machine learning data

To be of use for a machine learning algorithm when training a classifier, an annotated data set must be converted to training examples in which each training instance is described by the feature set of choice.

---

alternative sets of *given* coreferents, not to find the referents themselves.

Most machine learning algorithms require the presence of both positive and negative instances to be able to learn in an optimal manner. The type of features chosen determines the contents of the training instances; a training instance is often constituted by a pair of coreferring noun phrases (a positive example) or a pair of non-coreferring noun phrases (a negative example).

Perhaps the most intuitive, and naïve way of transforming an annotated corpus to training instances, would be to, for each coreferring pair of markables, consider the pair as a positive example, and all other pairs of markables as negative examples. However, such a transformation would mean that the resulting training data would be severely skewed: Ng and Cardie (2002a) point out that coreference is a rare relation, meaning that most noun phrases in a text are not coreferent, and that the standard MUC-6 and MUC-7 corpora contain only 2% positive instances. Ng and Cardie (2002a) propose a way of selecting training instances that samples both the negative and positive instances to build a more appropriate training set. The negative sample selection serves the purpose of reducing the amount of negative training data by selecting only those negative examples that lie between a coreferent markable and its farthest preceding antecedent. To remedy the drop in precision introduced by negative sample selection, Ng and Cardie propose a method for positive sample selection in order to find confident examples that are easy to learn. Ng and Cardie (2002a) show that their approach to sample selection improves overall system performance.

Harabagiu *et al* (2001) exploit coreference chains in order to explode the amount of training examples available from an annotated corpus. Harabagiu *et al* rely on the transitivity of coreference relations to obtain coreference chains. A coreference chain of length  $l$  can generate  $(l - 1)(l - 2)/2$  new coreference relations. Given, e.g., the 1845 original anaphoric relations available in the MUC-7 corpus, Harabagiu *et al* (2001) generate 15858 new such relations.

### 4.3. Weakly supervised approaches to reference resolution

A way of circumventing the problem of small data sets, is to exploit meta machine learning methods for bootstrapping a learner while, at the same time, building a larger data set.

Müller *et al* (2002) use co-training (Blum and Mitchell, 1998) in order to reduce the amount of manual work needed for creating training data. Co-training is a weakly-supervised meta learning algorithm that utilises two simple classifiers, a small amount of labeled training data, as well as a large pool unlabeled data, for bootstrapping a larger annotated data set. The two simple classifiers are first trained on different parts of the features describing the labeled training data. Then the co-training algorithm utilises the simple classifiers by letting them, in turn, label parts of

the unlabeled data. Instances labeled by one classifier is added to the other classifier’s training data. For co-training to work, the feature sets (*views*) used by the simple classifiers must be disjoint, but still elaborate enough for each classifier to learn the task at hand.

The results presented by Müller *et al* (2002) are “mostly negative”, and one source of the problem is that it is hard to split the feature set used in coreference resolution into the two disjoint and redundant views required by the co-training algorithm. Ng and Cardie (2003a) propose a single-view weakly-supervised algorithm to bootstrap classifiers without having to rely on a feature split. Their algorithm uses two different learning algorithms to train two classifiers using the same set of features on a split of the labeled training data. At each iteration, each classifier labels all the data in the pool of unlabeled data. The best instances, according to some metric, are added to the training data set of the other classifier, and vice versa. The pool of unlabeled data is re-populated with new data after each iteration. Ng and Cardie (2003a) show that their approach outperforms co-training in a comparative evaluation. Single-view training is claimed to be an alternative to co-training for tasks where no natural feature split has been found.

## 5. Evaluation

Three things can be said for sure about evaluation of reference resolution systems: There exists no perfect algorithm for resolving references, even in theory; The data sets that has become de-facto gold standard for evaluating reference resolution, MUC-6 and MUC-7 (see Section 4.1.), are claimed not to be perfect; The prevalent scheme for evaluating reference resolution is not uncontested.

Walker (1989) presents quantitative as well as qualitative results from a study of hand-simulating two recognized algorithms for pronoun resolution; that of Hobb’s (1986) and the centering approach presented by Brennan *et al* (1987). No significant difference is found between the performance of the two algorithms. Although results of algorithms cannot be compared out-of-context, it is worth mentioning that the results from Walker’s (1989) hand-simulated study place the accuracy of the Hobb’s algorithm and the centering approach in the high 80’s to low 90’s on the task of resolving pronouns.

van Deemter and Kibble (2000) argues that what is marked as coreference information in the MUC corpora are in fact not coreference relations proper; “As a result, it is not always clear what semantic relation these annotations are encoding.” Concluding the discussion about the annotation of reference, van Deemter and Kibble (2000) states “...we would like to submit that corpus-based research is sometimes insufficiently informed by theory.”

Further, the scheme for calculating the accuracy on the mentioned corpora (Vilain *et al.*, 1995) is

not uncontested. Bagga and Baldwin (1998) present the weaknesses and strengths for a number of different scoring algorithms, and come to the conclusion that the algorithm of Vilain *et al* suits applications that concern single coreference relations at a time rather than equivalence classes of coreferents. Bagga (1998) outlines a framework for evaluating coreference resolution systems.

Despite the above mentioned shortcomings, researchers are able to compare their efforts using the MUC data. The best performance by machine learning approaches to anaphora reference and coreference reported are on par with manually crafted systems, with F-scores ranging from mid 60's to lower 70's on MUC-6 and MUC-7 data, respectively.

## 6. Conclusions

Although the present survey of machine learning for anaphora and coreference does not cover all sources of information on the subject, some general observations can be made.

At first, hand-crafted reference systems were uncontested the best and the purpose of machine learning experiments on the subject were primarily to investigate the possibility to create a learning-system that could compete with a non-learning one. As learning systems gained in performance, the focus of the machine learning community shifted toward means to alleviate the need of annotated data. Today, learning systems are claimed to perform well on par with their non-learning siblings, and combinations of learning and non-learning systems seem to be the path of the future.

Two remarks to conclude this report: While the machine learning community seems to be focused on obtaining good performance using “cheap” methods, requiring as domain independent linguistic preprocessing as possible, the evaluation data of choice is still the rather narrow domain represented in the MUC-6 and MUC-7 corpora.

The argument by van Deemter and Kibble (2000) that the corpus-based research is insufficiently informed by theory seems valid and is very important; in general, the features used seem somewhat ad hoc and it is not always evident where they stem from.

## 7. References

- Aone, Chinatsu and Bennett, Scott William. 1995. Evaluation automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 122–129, Cambridge, MA, USA. ACL.
- Bagga, Amit and Baldwin, Breck. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at the First International Conference on Language Resources and Evaluation*, Granada, Spain, May.
- Bagga, Amit. 1998. Evaluation of coreferences and coreference resolution systems. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, May.
- Banko, Michele and Brill, Eric. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 26–33. ACL.
- Berger, Adam; Pietra, Stephen A. Della and Pietra, Vincent J. Della. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.
- Blum, Avrim and Mitchell, Tom. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, Madison, Wisconsin, USA, July. ACM.
- Boser, Bernhard E.; Guyon, Isabelle M. and Vapnik, Vladimir N. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 144–152. ACM.
- Brennan, Susan E.; Friedman, Marilyn Walker and Pollard, Carl J. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162, Stanford University, Stanford, CA, USA. ACL.
- Byron, Donna K. and Allen, James F. 1999. Applying genetic algorithms to pronoun resolution. In *Proceedings of the 16th National Conference on Artificial Intelligence*.
- Cardie, Claire and Wagstaff, Kiri. 1999. Noun phrase coreference as clustering. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 82–89. ACL.
- Chinchor, Nancy and Sundheim, Beth. 1996. Message understanding conference (muc) 6 additional news text. LDC96T10. FTP FILE. Philadelphia: Linguistic Data Consortium.
- Chinchor, Nancy and Sundheim, Beth. 2003. Message understanding conference (muc) 6. LDC2003T13. FTP FILE. Philadelphia: Linguistic Data Consortium.
- Cohen, William. 1995. Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*, Tahoe City, California, USA, July.
- Collins, Michael and Singer, Yoram. 1999. Unsupervised models for named entity classification. In *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. ACL, July.
- Connolly, Dennis; Burger, John D. and Day, David S. 1997. A machine learning approach to anaphoric ref-

- erence. In Jones, Daniel and Somers, Harold editors, *New Methods in Language Processing*, pages 133–144. UCL Press.
- Cortes, Corinna and Vapnik, Vladimir. 1995. Support-vector networks. *Machine Learning*, 22:1–25.
- Evans, Richard. 2001. Applying machine learning toward an automatic classification of it. *Literary and Linguistic Computing*, 16(1):45–57.
- Evans, Richard. 2002. Refined salience weighting and error analysis in anaphora resolution. In *Proceedings of Reference Resolution for Natural Language Processing*, pages 51–59, Alicante, Spain, June. University of Alicante.
- Ge, Niyu; Hale, John and Charniak, Eugene. 1998. A statistical approach to anaphora resolution. In *Proceedings of the 6th Workshop on Very Large Corpora*, pages 161–170.
- Grosz, Barbara J.; Joshi, Aravind K. and Weinstein, Scott. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.
- Halliday, Michael AK. and Hasan, Ruqaiya. 1976. *Cohesion in English*. Longman, London.
- Harabagiu, Sanda M.; Bunesco, Razvan C. and Maiorano, Steven J. 2001. Text and knowledge mining for coreference resolution. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics*, pages 55–62, Carnegie Mellon University, Pittsburg, PA, USA, June. ACL.
- Hirschman, Lynette and Chinchor, Nancy. 1997. Muc-7 coreference task definition (version 3.0). In *Proceedings of the 7th Message Understanding Conference (MUC-7)*.
- Hirschman, Lynette; Robinson, Patricia; Burger, John and Vilain, Marc. 1997. Automating coreference: the role of annotated training data. In *Proceedings of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*.
- Hobbs, Jerry R. 1986. Resolving pronoun references. In Spark-Jones, Karen and Webber, Bonnie editors, *Readings in Natural Language Processing*, pages 339–352. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Kehler, Andrew. 1997. Probabilistic coreference in information extraction. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, pages 163–173, Brown University, Providence, Rhode Island, USA, August. ACL.
- Lida, Ryu; Inui, Kentaro; Takamura, Hiroya and Matsumoto, Yuji. 2003. Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the Workshop on The Computational Treatment of Anaphora Resolution, held in conjunction with the 10th Conference of The European Chapter of the Association for Computational Linguistics*, pages 23–30, Budapest, Hungary, April. ACL.
- Linguistic Data Consortium, . 2001. Message understanding conference (muc) 7. LDC2001T02. FTP FILE. Philadelphia: Linguistic Data Consortium.
- Manning, Christopher D. and Schütze, Hinrich. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, Massachusetts, London, England.
- McCarthy, Joseph F. and Lehnert, Wendy G. 1995. Using decision trees for coreference resolution. In Mellish, C. editor, *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1050–1055.
- Mitchell, Tom. 1997. *Machine learning*. McGraw-Hill.
- Mitkov, Ruslan. 2001. Outstanding issues in anaphora resolution. In Gelbukh, Alexander F. editor, *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, Mexico-City, Mexico, February.
- Mitkov, Ruslan. 2002. *Anaphora resolution*. Studies in Language and Linguistics. Longman/Pearson Education, London, New York, Toronto, Sydney, Tokyo, Singapore, Hong Kong, Cape Town, New Delhi, Madrid, Paris, Amsterdam, Munich, Milan, Stockholm.
- Modejska, Natalia N.; Markert, Katja and Nissim, Malvina. 2003. Using the web in machine learning for other-anaphora resolution. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan, July. ACL.
- Müller, Christoph; Rapp, Stefan and Strube, Michael. 2002. Applying co-training to reference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 352–359, The Computer and Information Science Department and the Institute for Research in Cognitive Science, University of Pennsylvania Philadelphia, PA, USA, July. ACL.
- Ng, Vincent and Cardie, Claire. 2002a. Combining sample selection and error-driven pruning for machine learning of coreference rules. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 55–62, University of Pennsylvania, Philadelphia, PA, USA, July. ACL.
- Ng, Vincent and Cardie, Claire. 2002b. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics*, Howard International House, Taipei, Taiwan, August. ACL.
- Ng, Vincent and Cardie, Claire. 2002c. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, The Computer and Information Science Department and the Institute for Research in Cognitive Science, Univer-

- sity of Pennsylvania Philadelphia, PA, USA, July. ACL.
- Ng, Vincent and Cardie, Claire. 2003a. Bootstrapping coreference classifiers with multiple machine learning algorithms. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan, July. ACL.
- Ng, Vincent and Cardie, Claire. 2003b. Weakly supervised natural language learning without redundant views. In *Proceedings of Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting*, pages 173–180, Edmonton, Canada, May 27 - June 1. ACL.
- Nigam, Kamal; McCallum, Andrew; Thrun, Sebastian and Mitchell, Tom. 2000. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2/3):103–134.
- Preiss, Juditha. 2002. Anaphora resolution with memory based learning. In *Proceedings of the 5th Annual CLUK Research Colloquium*, pages 1–9, University of Leeds, UK, January. Computational Linguistics UK.
- Soderland, Stephen and Lehnert, Wendy G. 1994a. Corpus-driven knowledge acquisition for discourse analysis. In *Proceedings of the 12th National Conference on Artificial Intelligence*.
- Soderland, Stephen and Lehnert, Wendy G. 1994b. Wrap-up: a trainable discourse module for information extraction. *Journal of Artificial Intelligence Research*, 2:131–158.
- Soon, Wee Meng; Ng, Hwee Tou and Lim, Daniel Chung Yong. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, December.
- Strube, Michael; Rapp, Stefan and Müller, Christoph. 2002. The influence of minimum edit distance on reference resolution. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 312–319, University of Pennsylvania, Philadelphia, PA, USA, July. ACL.
- Stuckardt, Roland. 2002. Machine-learning-based vs. manually designed approaches to anaphor resolution: the best of two worlds. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium*, Lisbon, Portugal, September. University of Lisbon, Faculty of Sciences.
- Deemter, van Kees and Kibble, Rodger. 2000. On coreferring: coreference in muc and related annotation schemes. *Computational Linguistics*, 26(4):629–637, December.
- Vilain, Marc; Burger, John; Aberdeen, John; Connolly, Dennis and Hirschman, Lynette. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference*, pages 45–52, San Mateo, CA, USA. Morgan Kaufmann.
- Walker, Marily A. 1989. Evaluating discourse processing algorithms. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 251–261. ACL.
- Yang, Xiaofeng; Zhou, Guodong; Su, Jian and Tan, Chew Lim. 2003. Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 176–183. ACL, July.

Author(s)	Note on features	Data
Aone and Bennett (1995)	66 features; lexical, syntactic, semantic, positional (unary, binary). Features not listen in paper.	Japanese newspaper articles about joint ventures.
Connolly <i>et al</i> (1997)	15 features: type of anaphor and candidates; grammatical case of anaphora and candidates; distance from anaphor to each candidate; what candidate is more recent; agreement in count and gender; anaphor subsumes meaning of candidate	80 new-agency articles in English
Ge <i>et al</i> (1998)	4 features; distance, syntax, word information about proposed referent (gender, number, animacy), mention count.	Wall Street Journal
Byron and Allen (1999)	8 salience factors are subject to weighting by a genetic algorithm. Factors briefly outlined in the paper.	Penn Treebank
Evans (2001)	35 features grouped into six classes; positional, number of suggestional elements in surrounding text, lemmas in context, part of speech of a eight token surrounding window, indication of pleonastic use of <i>it</i> , proximity of following elements.	SUSANNE, BNC
Strube <i>et al</i> (2002)	15 features divided on document level (1 feature), noun phrase level (8 features), and coreference level (6 features): document id; grammatical function of antecedent; form of antecedent; agreement in person gender number of antecedent; semantic class of antecedent; grammatical function of anaphor; form of anaphor; agreement of person, gender, and number of anaphor; semantic class of anaphor; distance in words; distance in sentences; distance in markables; same grammatical function of anaphor and antecedent; anaphor and antecedent identical strings; substrings.	Texts in German about historical events.
Preiss (2002)	13 features; current sentence, subject, existential construct, possessive, direct object, indirect object, oblique, non embedded, non adjunct, current context, parallelism, locality, cataphora.	BNC
Evans (2002)	14 salience factors used by the non-learning system employed in this experiment are subject to weight assignment by a genetic algorithm.	English, computer hardware and software technical manuals, annual report from Amnesty International.
Stuckardt (2002)	Features not made explicit in the paper. Hybrid approach to resolution of non-possessive and possessive pronominal anaphora.	66 news agency press releases.
Modejska <i>et al</i> (2003)	2 different feature sets containing 9 and 11 features. The first set consists of: surface form; substring; grammatical role; anaphor-antecedent grammatical agreement; distance in sentences; semantic class; anaphor-antecedent gender agreement; type of relation. The second set consists of the same features as the first one, with the addition lexical knowledge from WordNet and information about named entities. Resolution of other-anaphora; 500 other-anaphors with antecedents.	Wall Street Journal (Penn Treebank)

Table 2: Overview of the features and data used for anaphora resolution in the surveyed papers. The features are elaborated on in respective paper, unless otherwise noted.

Author(s)	Notes on features	Data
Soderland and Lehnert (1994a; 1994b)	Features are automatically derived and dependent on the type of input. The system assumes that the input text has been processed so as to identify coreferences prior to invocation of the described module; the automatic derivation of features stems from combination of features available in the output from up-stream processing components. The task is Inter-sentential inference generation.	MUC-5
McCarthy and Lehnert (1995)	8 features; 2 on first reference (name, reference to "joint venture child"), 2 on second reference (same as for first reference), 4 on the pair of references (alias, pair is "joint venture child", share a common noun phrase, in same sentence).	MUC 5
Kehler (1997)	3 characteristics; related to templates filled by existing information extraction system, type of reference, and distance. Hybrid approach; assign probability distributions to alternative sets of coreferents.	No information available.
Cardie and Wagstaff (1999)	11 features describing each noun phrase; individual words, head noun, position, pronoun type, article, appositive, number, proper name, semantic class, gender, animacy.	MUC-6
Harabagiu <i>et al</i> (2001)	Indicators of: cohesion; gender, number, and class agreements; semantic consistency. The aim of Harabagiu <i>et al</i> 's (2001) approach is to discover minimalist coreference resolution rules, and as such, it is an effort in exploiting as few explicit characteristics of the data as possible.	MUC-6, MUC-7
Soon <i>et al</i> (2001)	12 features; distance, pronoun, antecedent pronoun, string match, definite noun phrase, demonstrative noun phrase, number agreement, semantic class agreement, gender agreement, both-proper-names, alias, appositive.	MUC-6, MUC-7
Ng and Cardie (2002c)	Experiments with two sets of features: the features of Soon <i>et al</i> (2001) combined with 41 additional features; and a reduced subset of 22-26 hand-picked features. The larger feature set caused Ng and Cardie's (2002c) to perform worse than did the smaller set.	MUC-6, MUC-7
Ng and Cardie (2002b)	Two tasks described; anaphoricity classification and coreference resolution. The anaphoricity classifier uses 37 features (elaborated on in the paper) grouped into four classes; lexical, grammatical, semantic, positional. The coreference classifier use the features outlined Ng and Cardie (2002c).	MUC-6, MUC-7

Table 3: Overview of the features and data used for coreference resolution in the surveyed papers. (*Continued on next page.*)

Author(s)	Notes on features	Data
Ng and Cardie (2002a)	25 features divided into five groups; lexical, grammatical, semantic, positional, others.	MUC-6, MUC-7
Müller <i>et al</i> (2002)	17 features at document level (1 feature), noun phrase level (8 features), and coreference level (8 features). The features are the same as reported by Strube <i>et al</i> (2002) with the addition of the features; minimum edit distance to anaphor, and minimum edit distance to antecedent.	Texts in German about historical events.
Ng and Cardie (2003a; 2003b)	25 features, see (Ng and Cardie, 2002a)	MUC-6, MUC-7
Lida <i>et al</i> (2003)	19 features divided into five groups; grammatical, semantic, positional, heuristic, centering. Lida <i>et al</i> (2003) adds features to capture the notion of Centering (Grosz <i>et al.</i> , 1995)	Japanese newspaper articles
Yang <i>et al</i> (2003)	23 features divided into four groups describing: the candidate for coreference; the anaphor; the anaphor and the candidate; the two candidates.	MUC-6, MUC-7

Table 3: (Continued from previous page.) Overview of the features and data used for coreference resolution in the surveyed papers.