Proceedings of the 48th Annual Meeting of the Association for
Computational Linguistics (ACL), Uppsala, Sweden, July 2010, pp. 1396--1411.

Supervised Noun Phrase Coreference Research: The First Fifteen Years

Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
vince@hlt.utdallas.edu

Abstract

The research focus of computational coreference resolution has exhibited a shift from heuristic approaches to machine learning approaches in the past decade. This paper surveys the major milestones in supervised coreference research since its inception fifteen years ago.

1 Introduction

Noun phrase (NP) coreference resolution, the task of determining which NPs in a text or dialogue refer to the same real-world entity, has been at the core of natural language processing (NLP) since the 1960s. NP coreference is related to the task of anaphora resolution, whose goal is to identify an antecedent for an *anaphoric* NP (i.e., an NP that depends on another NP, specifically its antecedent, for its interpretation) [see van Deemter and Kibble (2000) for a detailed discussion of the difference between the two tasks]. Despite its simple task definition, coreference is generally considered a difficult NLP task, typically involving the use of sophisticated knowledge sources and inference procedures (Charniak, 1972). Computational theories of discourse, in particular *focusing* (see Grosz (1977) and Sidner (1979)) and *centering* (Grosz et al. (1983; 1995)), have heavily influenced coreference research in the 1970s and 1980s, leading to the development of numerous *centering algorithms* (see Walker et al. (1998)).

The focus of coreference research underwent a gradual shift from heuristic approaches to machine learning approaches in the 1990s. This shift can be attributed in part to the advent of the statistical NLP era, and in part to the public availability of annotated coreference corpora produced as part of the MUC-6 (1995) and MUC-7 (1998) conferences. Learning-based coreference research has remained vibrant since then, with results regularly published not only in general NLP conferences, but also in specialized conferences (e.g., the biennial Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)) and workshops (e.g., the series of Bergen Workshop on Anaphora Resolution (WAR)). Being inherently a *clustering* task, coreference has also received a lot of attention in the machine learning community.

Fifteen years have passed since the first paper on learning-based coreference resolution was published (Connolly et al., 1994). Our goal in this paper is to provide NLP researchers with a survey of the major milestones in *supervised* coreference research, focusing on the computational models, the linguistic features, the annotated corpora, and the evaluation metrics that were developed in the past fifteen years. Note that several leading coreference researchers have published books (e.g., Mitkov (2002)), written survey articles (e.g., Mitkov (1999), Strube (2009)), and delivered tutorials (e.g., Strube (2002), Ponzetto and Poesio (2009)) that provide a broad overview of coreference research. This survey paper aims to *complement*, rather than supersede, these previously published materials. In particular, while existing survey papers discuss learning-based coreference research primarily in the context of the influential mention-pair model, we additionally survey recently proposed learning-based coreference models, which attempt to address the weaknesses of the mention-pair model. Due to space limitations, however, we will restrict our discussion to the most commonly investigated kind of coreference relation: the *identity* relation for NPs, excluding coreference among clauses and bridging references (e.g., part/whole and set/subset relations).

2 Annotated Corpora

The widespread popularity of machine learning approaches to coreference resolution can be attributed in part to the public availability of annotated coreference corpora. The MUC-6 and MUC-7 corpora, though relatively small (60 documents each) and homogeneous w.r.t. document type (newswire articles only), have been extensively used for training and evaluating coreference models. Equally popular are the corpora produced by the Automatic Content Extraction (ACE[1]) evaluations in the past decade: while the earlier ACE corpora (e.g., ACE-2) consist of solely English newswire and broadcast news articles, the later ones (e.g., ACE 2005) have also included Chinese and Arabic documents taken from additional sources such as broadcast conversations, webblog, usenet, and conversational telephone speech.

Coreference annotations are also publicly available in treebanks. These include (1) the English Penn Treebank (Marcus et al., 1993), which is labeled with coreference links as part of the OntoNotes project (Hovy et al., 2006); (2) the Tübingen Treebank (Telljohann et al., 2004), which is a collection of German news articles consisting of 27,125 sentences; (3) the Prague Dependency Treebank (Hajič et al., 2006), which consists of 3168 news articles taken from the Czech National Corpus; (4) the NAIST Text Corpus (Iida et al., 2007b), which consists of 287 Japanese news articles; (5) the AnCora Corpus (Recasens and Martí, 2009), which consists of Spanish and Catalan journalist texts; and (6) the GENIA corpus (Ohta et al., 2002), which contains 2000 MEDLINE abstracts.

Other publicly available coreference corpora of interest include two annotated by Ruslan Mitkov's research group: (1) a 55,000-word corpus in the domain of security/terrorism (Hasler et al., 2006); and (2) training data released as part of the 2007 Anaphora Resolution Exercise (Orăsan et al., 2008), a coreference resolution shared task. There are also two that consist of spoken dialogues: the TRAINS93 corpus (Heeman and Allen, 1995) and the Switchboard data set (Calhoun et al., in press).

Additional coreference data will be available in the near future. For instance, the SemEval-2010 shared task on Coreference Resolution in Multiple Languages (Recasens et al., 2009) has promised to release coreference data in six languages. In addition, Massimo Poesio and his colleagues are leading an annotation project that aims to collect large amounts of coreference data for English via a Web Collaboration game called Phrase Detectives[2].

[1] http://www.itl.nist.gov/iad/mig/tests/ace/

[2] http://www.phrasedetectives.org

3 Learning-Based Coreference Models

In this section, we examine three important classes of coreference models that were developed in the past fifteen years, namely, the mention-pair model, the entity-mention model, and ranking models.

3.1 Mention-Pair Model

The mention-pair model is a classifier that determines whether two NPs are coreferent. It was first proposed by Aone and Bennett (1995) and McCarthy and Lehnert (1995), and is one of the most influential learning-based coreference models. Despite its popularity, this binary classification approach to coreference is somewhat undesirable: the transitivity property inherent in the coreference relation cannot be enforced, as it is possible for the model to determine that A and B are coreferent, B and C are coreferent, but A and C are not coreferent. Hence, a separate clustering mechanism is needed to coordinate the pairwise classification decisions made by the model and construct a coreference partition.

Another issue that surrounds the acquisition of the mention-pair model concerns the way training instances are created. Specifically, to determine whether a pair of NPs is coreferent or not, the mention-pair model needs to be trained on a data set where each instance represents two NPs and possesses a class value that indicates whether the two NPs are coreferent. Hence, a natural way to assemble a training set is to create one instance from each pair of NPs appearing in a training document. However, this instance creation method is rarely employed: as most NP pairs in a text are not coreferent, this method yields a training set with a skewed class distribution, where the negative instances significantly outnumber the positives.

As a result, in practical implementations of the mention-pair model, one needs to specify not only the *learning algorithm* for training the model and the *linguistic features* for representing an instance, but also the *training instance creation method* for reducing class skewness and the *clustering algorithm* for constructing a coreference partition.

3.1.1 Creating Training Instances

As noted above, the primary purpose of training instance creation is to reduce class skewness. Many heuristic instance creation methods have been proposed, among which Soon et al.'s (1999; 2001) is arguably the most popular choice. Given an anaphoric noun phrase[3], NP_k, Soon et al.'s method creates a *positive instance* between NP_k and its closest preceding antecedent, NP_j, and a *negative instance* by pairing NP_k with each of the intervening NPs, $\text{NP}_{j+1}, \ldots, \text{NP}_{k-1}$.

With an eye towards improving the precision of a coreference resolver, Ng and Cardie (2002c) propose an instance creation method that involves a single modification to Soon et al.'s method: if NP_k is non-pronominal, a positive instance should be formed between NP_k and its closest preceding *non-pronominal* antecedent instead. This modification is motivated by the observation that it is not easy for a human, let alone a machine learner, to learn from a positive instance where the antecedent of a non-pronominal NP is a pronoun.

To further reduce class skewness, some researchers employ a filtering mechanism on top of an instance creation method, thereby disallowing the creation of training instances from NP pairs that are unlikely to be coreferent, such as NP pairs that violate gender and number agreement (e.g., Strube et al. (2002), Yang et al. (2003)).

While many instance creation methods are heuristic in nature (see Uryupina (2004) and Hoste and Daelemans (2005)), some are learning-based. For example, motivated by the fact that some coreference relations are harder to identify than the others (see Harabagiu et al. (2001)), Ng and Cardie (2002a) present a method for mining easy positive instances, in an attempt to avoid the inclusion of hard training instances that may complicate the acquisition of an accurate coreference model.

3.1.2 Training a Coreference Classifier

Once a training set is created, we can train a coreference model using an off-the-shelf learning algorithm. Decision tree induction systems (e.g., C5 (Quinlan, 1993)) are the first and one of the most widely used learning algorithms by coreference researchers, although rule learners (e.g., RIPPER (Cohen, 1995)) and memory-based learners (e.g., TiMBL (Daelemans and Van den Bosch, 2005)) are also popular choices, especially in early applications of machine learning to coreference resolution. In recent years, statistical learners such as maximum entropy models (Berger et al., 1996), voted perceptrons (Freund and Schapire, 1999), and support vector machines (Joachims, 1999) have been increasingly used, in part due to their ability to provide a confidence value (e.g., in the form of a probability) associated with a classification, and in part due to the fact that they can be easily adapted to train recently proposed ranking-based coreference models (see Section 3.3).

3.1.3 Generating an NP Partition

After training, we can apply the resulting model to a test text, using a clustering algorithm to coordinate the pairwise classification decisions and impose an NP partition. Below we describe some commonly used coreference clustering algorithms.

Despite their simplicity, *closest-first clustering* (Soon et al., 2001) and *best-first clustering* (Ng and Cardie, 2002c) are arguably the most widely used coreference clustering algorithms. The closest-first clustering algorithm selects as the antecedent for an NP, NP_k, the closest preceding noun phrase that is classified as coreferent with it.[4] However, if no such preceding noun phrase exists, no antecedent is selected for NP_k. The *best-first* clustering algorithm aims to improve the precision of closest-first clustering, specifically by selecting as the antecedent of NP_k the *most probable* preceding NP that is classified as coreferent with it.

One criticism of the closest-first and best-first clustering algorithms is that they are too greedy. In particular, clusters are formed based on a small subset of the pairwise decisions made by the model. Moreover, positive pairwise decisions are unjustifiably favored over their negative counterparts. For example, three NPs are likely to end up in the same cluster in the resulting partition even if there is strong evidence that A and C are not coreferent, as long as the other two pairs (i.e., (A,B) and (B,C)) are classified as positive.

Several algorithms that address one or both of these problems have been used for coreference clustering. *Correlation clustering* (Bansal et al., 2002), which produces a partition that respects as many pairwise decisions as possible, is used by McCallum and Wellner (2004), Zelenko et al. (2004), and Finley and Joachims (2005). *Graph partitioning algorithms* are applied on a weighted, undirected graph where a vertex corresponds to an NP and an edge is weighted by the pairwise coreference scores between two NPs (e.g., McCallum and Wellner (2004), Nicolae and Nico-

[3] In this paper, we use the term *anaphoric* to describe any NP that is part of a coreference chain but is not the head of the chain. Hence, proper names can be anaphoric under this overloaded definition, but linguistically, they are not.

[4] If a probabilistic model is used, we can define a threshold above which a pair of NPs is considered coreferent.

lae (2006)). The *Dempster-Shafer rule* (Dempster, 1968), which combines the positive and negative pairwise decisions to score a partition, is used by Kehler (1997) and Bean and Riloff (2004) to identify the most probable NP partition.

Some clustering algorithms bear a closer resemblance to the way a human creates coreference clusters. In these algorithms, not only are the NPs in a text processed in a left-to-right manner, the later coreference decisions are dependent on the earlier ones (Cardie and Wagstaff, 1999; Klenner and Ailloud, 2008).[5] For example, to resolve an NP, NP_k, Cardie and Wagstaff's algorithm considers each preceding NP, NP_j, as a candidate antecedent in a right-to-left order. If NP_k and NP_j are likely to be coreferent, the algorithm imposes an additional check that NP_k does not violate any constraint on coreference (e.g., gender agreement) with any NP in the cluster containing NP_j before positing that the two NPs are coreferent.

Luo et al.'s (2004) Bell-tree-based algorithm is another clustering algorithm where the later coreference decisions are dependent on the earlier ones. A Bell tree provides an elegant way of organizing the space of NP partitions. Informally, a node in the ith level of a Bell tree corresponds to an ith-order *partial* partition (i.e., a partition of the first i NPs of the given document), and the ith level of the tree contains *all* possible ith-order partial partitions. Hence, a leaf node contains a *complete* partition of the NPs, and the goal is to search for the leaf node that contains the most probable partition. The search starts at the root, and a partitioning of the NPs is incrementally constructed as we move down the tree. Specifically, based on the coreference decisions it has made in the first $i-1$ levels of the tree, the algorithm determines at the ith level whether the ith NP should start a new cluster, or to which preceding *cluster* it should be assigned.

While many coreference clustering algorithms have been developed, there have only been a few attempts to compare their effectiveness. For example, Ng and Cardie (2002c) report that best-first clustering is better than closest-first clustering. Nicolae and Nicolae (2006) show that best-first clustering performs similarly to Bell-tree-based clustering, but neither of these algorithms performs as well as their proposed minimum-cut-based graph partitioning algorithm.

3.1.4 Determining NP Anaphoricity

While coreference clustering algorithms attempt to resolve *each* NP encountered in a document, only a subset of the NPs are *anaphoric* and therefore need to be resolved. Hence, knowledge of the anaphoricity of an NP can potentially improve the precision of a coreference resolver.

Traditionally, the task of anaphoricity determination has been tackled independently of coreference resolution using a variety of techniques. For example, pleonastic *it* has been identified using heuristic approaches (e.g., Paice and Husk (1987), Lappin and Leass (1994), Kennedy and Boguraev (1996)), supervised approaches (e.g., Evans (2001), Müller (2006), Versley et al. (2008a)), and distributional methods (e.g., Bergsma et al. (2008)); and non-anaphoric definite descriptions have been identified using rule-based techniques (e.g., Vieira and Poesio (2000)) and unsupervised techniques (e.g., Bean and Riloff (1999)).

Recently, anaphoricity determination has been evaluated in the context of coreference resolution, with results showing that training an anaphoricity classifier to identify and filter non-anaphoric NPs prior to coreference resolution can improve a learning-based resolver (e.g., Ng and Cardie (2002b), Uryupina (2003), Poesio et al. (2004b)). Compared to earlier work on anaphoricity determination, recently proposed approaches are more "global" in nature, taking into account the pairwise decisions made by the mention-pair model when making anaphoricity decisions. Examples of such approaches have exploited techniques including integer linear programming (ILP) (Denis and Baldridge, 2007a), label propagation (Zhou and Kong, 2009), and minimum cuts (Ng, 2009).

3.1.5 Combining Classification & Clustering

From a learning perspective, a two-step approach to coreference — classification and clustering — is undesirable. Since the classification model is trained independently of the clustering algorithm, improvements in classification accuracy do not guarantee corresponding improvements in clustering-level accuracy. That is, overall performance on the coreference task might not improve.

To address this problem, McCallum and Wellner (2004) and Finley and Joachims (2005) eliminate the classification step entirely, treating coref-

[5] When applying closest-first and best-first clustering, Soon et al. (2001) and Ng and Cardie (2002c) also process the NPs in a sequential manner, but since the later decisions are not dependent on the earlier ones, the order in which the NPs are processed does not affect their clustering results.

erence as a *supervised clustering* task where a similarity metric is learned to directly maximize clustering accuracy. Klenner (2007) and Finkel and Manning (2008) use ILP to ensure that the pairwise classification decisions satisfy transitivity.[6]

3.1.6 Weaknesses of the Mention-Pair Model

While many of the aforementioned algorithms for clustering and anaphoricity determination have been shown to improve coreference performance, the underlying model with which they are used in combination — the mention-pair model — remains fundamentally weak. The model has two commonly-cited weaknesses. First, since each candidate antecedent for an anaphoric NP to be resolved is considered independently of the others, the model only determines how good a candidate antecedent is relative to the anaphoric NP, but not how good a candidate antecedent is relative to other candidates. In other words, it fails to answer the question of which candidate antecedent is most probable. Second, it has limitations in its expressiveness: the information extracted from the two NPs alone may not be sufficient for making an informed coreference decision, especially if the candidate antecedent is a pronoun (which is semantically empty) or a mention that lacks descriptive information such as gender (e.g., "Clinton"). Below we discuss how these weaknesses are addressed by the entity-mention model and ranking models.

3.2 Entity-Mention Model

The entity-mention model addresses the expressiveness problem with the mention-pair model. To motivate the entity-mention model, consider an example taken from McCallum and Wellner (2003), where a document consists of three NPs: "Mr. Clinton," "Clinton," and "she." The mention-pair model may determine that "Mr. Clinton" and "Clinton" are coreferent using string-matching features, and that "Clinton" and "she" are coreferent based on proximity and lack of evidence for gender and number disagreement. However, these two pairwise decisions together with transitivity imply that "Mr. Clinton" and "she" will end up in the same cluster, which is incorrect due to gender mismatch. This kind of error arises in part because the later coreference decisions are not dependent on the earlier ones. In particular, had the model taken into consideration that "Mr. Clinton" and "Clinton" were in the same cluster, it probably would not have posited that "she" and "Clinton" are coreferent. The aforementioned Cardie and Wagstaff algorithm attempts to address this problem in a *heuristic* manner. It would be desirable to *learn* a model that can classify whether an NP to be resolved is coreferent with a preceding, possibly partially-formed, cluster. This model is commonly known as the entity-mention model.

Since the entity-mention model aims to classify whether an NP is coreferent with a preceding cluster, each of its training instances (1) corresponds to an NP, NP_k, and a preceding cluster, C_j, and (2) is labeled with either POSITIVE or NEGATIVE, depending on whether NP_k should be assigned to C_j. Consequently, we can represent each instance by a set of *cluster-level* features (i.e., features that are defined over an arbitrary subset of the NPs in C_j). A cluster-level feature can be computed from a feature employed by the mention-pair model by applying a logical predicate. For example, given the NUMBER AGREEMENT feature, which determines whether two NPs agree in number, we can apply the ALL predicate to create a cluster-level feature, which has the value YES if NP_k agrees in number with *all* of the NPs in C_j and NO otherwise. Other commonly-used logical predicates for creating cluster-level features include relaxed versions of the ALL predicate, such as MOST, which is true if NP_k agrees in number with more than half of the NPs in C_j, and ANY, which is true as long as NP_k agrees in number with just one of the NPs in C_j. The ability of the entity-mention model to employ cluster-level features makes it more expressive than its mention-pair counterpart.

Despite its improved expressiveness, the entity-mention model has not yielded particularly encouraging results. For example, Luo et al. (2004) apply the ANY predicate to generate cluster-level features for their entity-mention model, which does not perform as well as the mention-pair model. Yang et al. (2004b; 2008a) also investigate the entity-mention model, which produces results that are only marginally better than those of the mention-pair model. However, it appears that they are not fully exploiting the expressiveness of the entity-mention model, as cluster-level features only comprise a small fraction of their features.

Variants of the entity-mention model have been investigated. For example, Culotta et al. (2007) present a first-order logic model that determines

[6]Recently, however, Klenner and Ailloud (2009) have become less optimistic about ILP approaches to coreference.

the probability that an arbitrary set of NPs are all co-referring. Their model resembles the entity-mention model in that it enables the use of cluster-level features. Daumé III and Marcu (2005) propose an online learning model for constructing coreference chains in an incremental fashion, allowing later coreference decisions to be made by exploiting cluster-level features that are computed over the coreference chains created thus far.

3.3 Ranking Models

While the entity-mention model addresses the expressiveness problem with the mention-pair model, it does not address the other problem: failure to identify the most probable candidate antecedent. Ranking models, on the other hand, allow us to determine which candidate antecedent is most probable given an NP to be resolved. Ranking is arguably a more natural reformulation of coreference resolution than classification, as a ranker allows all candidate antecedents to be considered *simultaneously* and therefore directly captures the competition among them. Another desirable consequence is that there exists a natural resolution strategy for a ranking approach: an anaphoric NP is resolved to the candidate antecedent that has the highest rank. This contrasts with classification-based approaches, where many clustering algorithms have been employed to coordinate the pairwise classification decisions, and it is still not clear which of them is the best.

The notion of ranking candidate antecedents can be traced back to centering algorithms, many of which use grammatical roles to rank forward-looking centers (see Walker et al. (1998)). Ranking is first applied to learning-based coreference resolution by Connolly et al. (1994; 1997), where a model is trained to rank two candidate antecedents. Each training instance corresponds to the NP to be resolved, NP_k, as well as two candidate antecedents, NP_i and NP_j, one of which is an antecedent of NP_k and the other is not. Its class value indicates which of the two candidates is better. This model is referred to as the *tournament* model by Iida et al. (2003) and the *twin-candidate* model by Yang et al. (2003; 2008b). To resolve an NP during testing, one way is to apply the model to each pair of its candidate antecedents, and the candidate that is classified as better the largest number of times is selected as its antecedent.

Advances in machine learning have made it possible to train a *mention ranker* that ranks *all* of the candidate antecedents simultaneously. While mention rankers have consistently outperformed the mention-pair model (Versley, 2006; Denis and Baldridge, 2007b), they are not more expressive than the mention-pair model, as they are unable to exploit cluster-level features, unlike the entity-mention model. To enable rankers to employ cluster-level features, Rahman and Ng (2009) propose the cluster-ranking model, which ranks preceding *clusters*, rather than candidate antecedents, for an NP to be resolved. Cluster rankers therefore address both weaknesses of the mention-pair model, and have been shown to improve mention rankers. Cluster rankers are conceptually similar to Lappin and Leass's (1994) heuristic pronoun resolver, which resolves an anaphoric pronoun to the most salient preceding cluster.

An important issue with ranking models that we have eluded so far concerns the identification of non-anaphoric NPs. As a ranker simply imposes a ranking on candidate antecedents or preceding clusters, it cannot determine whether an NP is anaphoric (and hence should be resolved). To address this problem, Denis and Baldridge (2008) apply an independently trained anaphoricity classifier to identify non-anaphoric NPs prior to ranking, and Rahman and Ng (2009) propose a model that jointly learns coreference and anaphoricity.

4 Knowledge Sources

Another thread of supervised coreference research concerns the development of linguistic features. Below we give an overview of these features.

String-matching features can be computed robustly and typically contribute a lot to the performance of a coreference system. Besides simple string-matching operations such as exact string match, substring match, and head noun match for different kinds of NPs (see Daumé III and Marcu (2005)), slightly more sophisticated string-matching facilities have been attempted, including minimum edit distance (Strube et al., 2002) and longest common subsequence (Castaño et al., 2002). Yang et al. (2004a) treat the two NPs involved as two bags of words, and compute their similarity using metrics commonly-used in information retrieval, such as the dot product, with each word weighted by their TF-IDF value.

Syntactic features are computed based on a syntactic parse tree. Ge et al. (1998) implement a *Hobbs distance* feature, which encodes the rank assigned to a candidate antecedent for a pronoun by Hobbs's (1978) seminal syntax-based pronoun resolution algorithm. Luo and Zitouni (2005) extract features from a parse tree for implementing Binding Constraints (Chomsky, 1988). Given an automatically parsed corpus, Bergsma and Lin (2006) extract from each parse tree a dependency path, which is represented as a sequence of nodes and dependency labels connecting a pronoun and a candidate antecedent, and collect statistical information from these paths to determine the likelihood that a pronoun and a candidate antecedent connected by a given path are coreferent. Rather than deriving features from parse trees, Iida et al. (2006) and Yang et al. (2006) employ these trees directly as *structured* features for pronoun resolution. Specifically, Yang et al. define tree kernels for efficiently computing the similarity between two parse trees, and Iida et al. use a boosting-based algorithm to compute the usefulness of a subtree.

Grammatical features encode the grammatical properties of one or both NPs involved in an instance. For example, Ng and Cardie's (2002c) resolver employs 34 grammatical features. Some features determine NP type (e.g., are both NPs definite or pronouns?). Some determine the grammatical role of one or both of the NPs. Some encode traditional linguistic (hard) constraints on coreference. For example, coreferent NPs have to agree in number and gender and cannot span one another (e.g., "Google" and "Google employees"). There are also features that encode general linguistic preferences either for or against coreference. For example, an indefinite NP (that is not in apposition to an anaphoric NP) is not likely to be coreferent with any NP that precedes it.

There has been an increasing amount of work on investigating **semantic features** for coreference resolution. One of the earliest kinds of semantic knowledge employed for coreference resolution is perhaps selectional preference (Dagan and Itai, 1990; Kehler et al., 2004b; Yang et al., 2005; Haghighi and Klein, 2009): given a pronoun to be resolved, its governing verb, and its grammatical role, we prefer a candidate antecedent that can be governed by the same verb and be in the same role. Semantic knowledge has also been extracted from WordNet and unannotated corpora for computing the semantic compatibility/similarity between two common nouns (Harabagiu et al., 2001; Versley, 2007) as well as the semantic class of a noun (Ng, 2007a; Huang et al., 2009). One difficulty with deriving knowledge from WordNet is that one has to determine which sense of a given word to use. Some researchers simply use the first sense (Soon et al., 2001) or all possible senses (Ponzetto and Strube, 2006a), while others overcome this problem with word sense disambiguation (Nicolae and Nicolae, 2006). Knowledge has also been mined from Wikipedia for measuring the semantic relatedness of two NPs, NP_j and NP_k (Ponzetto and Strube (2006a; 2007)), such as: whether $\text{NP}_{j/k}$ appears in the first paragraph of the Wiki page that has $\text{NP}_{k/j}$ as the title or in the list of categories to which this page belongs, and the degree of overlap between the two pages that have the two NPs as their titles (see Poesio et al. (2007) for other uses of encyclopedic knowledge for coreference resolution). Contextual roles (Bean and Riloff, 2004), semantic relations (Ji et al., 2005), semantic roles (Ponzetto and Strube, 2006b; Kong et al., 2009), and animacy (Orăsan and Evans, 2007) have also been exploited to improve coreference resolution.

Lexico-syntactic *patterns* have been used to capture the semantic relatedness between two NPs and hence the likelihood that they are coreferent. For instance, given the pattern *X is a Y* (which is highly indicative that *X* and *Y* are coreferent), we can instantiate it with a pair of NPs and search for the instantiated pattern in a large corpus or the Web (Daumé III and Marcu, 2005; Haghighi and Klein, 2009). The more frequently the pattern occurs, the more likely they are coreferent. This technique has been applied to resolve different kinds of anaphoric references, including *other*-anaphora (Modjeska et al., 2003; Markert and Nissim, 2005) and bridging references (Poesio et al., 2004a). While these patterns are typically hand-crafted (e.g., Garera and Yarowsky (2006)), they can also be learned from an annotated corpus (Yang and Su, 2007) or bootstrapped from an unannotated corpus (Bean and Riloff, 2004).

Despite the large amount of work on discourse-based anaphora resolution in the 1970s and 1980s (see Hirst (1981)), learning-based resolvers have only exploited shallow **discourse-based features**, which primarily involve characterizing the salience of a candidate antecedent by measuring its distance from the anaphoric NP to be resolved or determining whether it is in a prominent grammatical role (e.g., subject). A notable exception is Iida et al. (2009), who train a ranker to rank the candidate antecedents for an anaphoric pronoun by their salience. It is worth noting that Tetreault (2005) has employed Grosz and Sidner's (1986) discourse theory and Veins Theory (Ide and Cristea, 2000) to identify and remove candidate antecedents that are not referentially accessible to an anaphoric pronoun in his heuristic pronoun resolvers. It would be interesting to incorporate this idea into a learning-based resolver.

There are also features that do not fall into any of the preceding categories. For example, a memorization feature is a word pair composed of the head nouns of the two NPs involved in an instance (Bengtson and Roth, 2008). Memorization features have been used as binary-valued features indicating the presence or absence of their words (Luo et al., 2004) or as probabilistic features indicating the probability that the two heads are coreferent according to the training data (Ng, 2007b). An anaphoricity feature indicates whether an NP to be resolved is anaphoric, and is typically computed using an anaphoricity classifier (Ng, 2004), hand-crafted patterns (Daumé III and Marcu, 2005), and automatically acquired patterns (Bean and Riloff, 1999). Finally, the outputs of rule-based pronoun and coreference resolvers have also been used as features for learning-based coreference resolution (Ng and Cardie, 2002c).

For an empirical evaluation of the contribution of a subset of these features to the mention-pair model, see Bengtson and Roth (2008).

5 Evaluation Issues

Two important issues surround the evaluation of a coreference resolver. First, how do we obtain the set of NPs that a resolver will partition? Second, how do we score the partition it produces?

5.1 Extracting Candidate Noun Phrases

To obtain the set of NPs to be partitioned by a resolver, three methods are typically used. In the first method, the NPs are extracted automatically from a syntactic parser. The second method involves extracting the NPs directly from the gold standard. In the third method, a *mention detector* is first trained on the gold-standard NPs in the training texts, and is then applied to automatically extract *system mentions* in a test text.[7] Note that these three extraction methods typically produce different numbers of NPs: the NPs extracted from a parser tend to significantly outnumber the system mentions, which in turn outnumber the gold NPs. The reasons are two-fold. First, in some coreference corpora (e.g., MUC-6 and MUC-7), the NPs that are not part of any coreference chain are not annotated. Second, in corpora such as those produced by the ACE evaluations, only the NPs that belong to one of the ACE entity types (e.g., PERSON, ORGANIZATION, LOCATION) are annotated.

Owing in large part to the difference in the number of NPs extracted by these three methods, a coreference resolver can produce substantially different results when applied to the resulting three sets of NPs, with gold NPs yielding the best results and NPs extracted from a parser yielding the worst (Nicolae and Nicolae, 2006). While researchers who evaluate their resolvers on gold NPs point out that the results can more accurately reflect the performance of their coreference algorithm, Stoyanov et al. (2009) argue that such evaluations are unrealistic, as NP extraction is an integral part of an end-to-end fully-automatic resolver.

Whichever NP extraction method is employed, it is clear that the use of gold NPs can considerably simplify the coreference task, and hence resolvers employing different extraction methods should *not* be compared against each other.

5.2 Scoring a Coreference Partition

The MUC scorer (Vilain et al., 1995) is the first program developed for scoring coreference partitions. It has two often-cited weaknesses. As a *link-based* measure, it does not reward correctly identified singleton clusters since there is no coreference link in these clusters. Also, it tends to underpenalize partitions with overly large clusters.

To address these problems, two coreference scoring programs have been developed: B^3 (Bagga and Baldwin, 1998) and CEAF (Luo, 2005). Note that both scorers have only been defined for the case where the key partition has the same set of NPs as the response partition. To apply these scorers to automatically extracted NPs, different methods have been proposed (see Rahman and Ng (2009) and Stoyanov et al. (2009)).

Since coreference is a clustering task, any general-purpose method for evaluating a response partition against a key partition (e.g., Kappa (Carletta, 1996)) can be used for coreference scor-

[7]An exception is Daumé III and Marcu (2005), whose model jointly learns to extract NPs and perform coreference.

ing (see Popescu-Belis et al. (2004)). In practice, these general-purpose methods are typically used to provide scores that complement those obtained via the three coreference scorers discussed above. It is worth mentioning that there is a trend towards evaluating a resolver against multiple scorers, which can indirectly help to counteract the bias inherent in a particular scorer. For further discussion on evaluation issues, see Byron (2001).

6 Concluding Remarks

While we have focused our discussion on supervised approaches, coreference researchers have also attempted to reduce a resolver's reliance on annotated data by combining a small amount of labeled data and a large amount of unlabeled data using general-purpose semi-supervised learning algorithms such as co-training (Müller et al., 2002), self-training (Kehler et al., 2004a), and EM (Cherry and Bergsma, 2005; Ng, 2008). Interestingly, recent results indicate that unsupervised approaches to coreference resolution (e.g., Haghighi and Klein (2007; 2010), Poon and Domingos (2008)) rival their supervised counterparts, casting doubts on whether supervised resolvers are making effective use of the available labeled data.

Another issue that we have not focused on but which is becoming increasingly important is multilinguality. While many of the techniques discussed in this paper were originally developed for English, they have been applied to learn coreference models for other languages, such as Chinese (e.g., Converse (2006)), Japanese (e.g., Iida (2007)), Arabic (e.g., Luo and Zitouni (2005)), Dutch (e.g., Hoste (2005)), German (e.g., Wunsch (2010)), Swedish (e.g., Nilsson (2010)), and Czech (e.g., Nguy et al. (2009)). In addition, researchers have developed approaches that are targeted at handling certain kinds of anaphora present in non-English languages, such as zero anaphora (e.g., Iida et al. (2007a), Zhao and Ng (2007)).

As Mitkov (2001) puts it, coreference resolution is a "difficult, but not intractable problem," and we have been making "slow, but steady progress" on improving machine learning approaches to the problem in the past fifteen years. To ensure further progress, researchers should compare their results against a baseline that is stronger than the commonly-used Soon et al. (2001) system, which relies on a weak model (i.e., the mention-pair model) and a small set of linguistic features. As recent systems are becoming more sophisticated, we suggest that researchers make their systems publicly available in order to facilitate performance comparisons. Publicly available coreference systems currently include JavaRAP (Qiu et al., 2004), GuiTaR (Poesio and Kabadjov, 2004), BART (Versley et al., 2008b), CoRTex (Denis and Baldridge, 2008), the Illinois Coreference Package (Bengtson and Roth, 2008), CherryPicker (Rahman and Ng, 2009), Reconcile (Stoyanov et al., 2010), and Charniak and Elsner's (2009) pronoun resolver.

We conclude with a discussion of two questions regarding supervised coreference research. First, *what is the state of the art?* This is not an easy question, as researchers have been evaluating their resolvers on different corpora using different evaluation metrics and preprocessing tools. In particular, preprocessing tools can have a large impact on the performance of a resolver (Barbu and Mitkov, 2001). Worse still, assumptions about whether gold or automatically extracted NPs are used are sometimes not explicitly stated, potentially causing results to be interpreted incorrectly. To our knowledge, however, the best results on the MUC-6 and MUC-7 data sets using automatically extracted NPs are reported by Yang et al. (2003) (71.3 MUC F-score) and Ng and Cardie (2002c) (63.4 MUC F-score), respectively;[8] and the best results on the ACE data sets using gold NPs can be found in Luo (2007) (88.4 ACE-value).

Second, *what lessons can we learn from fifteen years of learning-based coreference research?* The mention-pair model is weak because it makes coreference decisions based on local information (i.e., information extracted from two NPs). Expressive models (e.g., those that can exploit cluster-level features) generally offer better performance, and so are models that are "global" in nature. Global coreference models may refer to any kind of models that can exploit non-local information, including models that can consider multiple candidate antecedents simultaneously (e.g., ranking models), models that allow joint learning for coreference resolution and related tasks (e.g., anaphoricity determination), models that can directly optimize clustering-level (rather than classification) accuracy, and models that can coordinate with other components of a resolver, such as training instance creation and clustering.

[8] These results by no means suggest that no progress has been made since 2003: most of the recently proposed coreference models were evaluated on the ACE data sets.

Acknowledgments

We thank the three anonymous reviewers for their invaluable comments on an earlier draft of the paper. This work was supported in part by NSF Grant IIS-0812261. Any opinions, findings, and conclusions or recommendations expressed are those of the author and do not necessarily reflect the views or official policies, either expressed or implied, of the NSF.

References

Chinatsu Aone and Scott William Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 122–129.

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

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2002. Correlation clustering. In *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, pages 238–247.

Catalina Barbu and Ruslan Mitkov. 2001. Evaluation tool for rule-based anaphora resolution methods. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 34–41.

David Bean and Ellen Riloff. 1999. Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 373–380.

David Bean and Ellen Riloff. 2004. Unsupervised learning of contextual role knowledge for coreference resolution. In *Human Language Technologies 2004: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 297–304.

Eric Bengtson and Dan Roth. 2008. Understanding the values of features for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 294–303.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

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

Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Distributional identification of non-referential pronouns. In *Proceedings of ACL-08: HLT*, pages 10–18.

Donna Byron. 2001. The uncommon denominator: A proposal for consistent reporting of pronoun resolution results. *Computational Linguistics*, 27(4):569–578.

Sasha Calhoun, Jean Carletta, Jason Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. (in press). The NXT-format Switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*.

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

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

José Castaño, Jason Zhang, and James Pustejovsky. 2002. Anaphora resolution in biomedical literature. In *Proceedings of the 2002 International Symposium on Reference Resolution*.

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*, pages 148–156.

Eugene Charniak. 1972. *Towards a Model of Children's Story Comphrension*. AI-TR 266, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA.

Colin Cherry and Shane Bergsma. 2005. An expectation maximization approach to pronoun resolution. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 88–95.

Noam Chomsky. 1988. *Language and Problems of Knowledge. The Managua Lectures*. MIT Press, Cambridge, Massachusetts.

William Cohen. 1995. Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*, pages 115–123.

Dennis Connolly, John D. Burger, and David S. Day. 1994. A machine learning approach to anaphoric reference. In *Proceedings of International Conference on New Methods in Language Processing*, pages 255–261.

Dennis Connolly, John D. Burger, and David S. Day. 1997. A machine learning approach to anaphoric reference. In D. Jones and H. Somers, editors, *New Methods in Language Processing*, pages 133–144. UCL Press.

Susan Converse. 2006. *Pronominal Anaphora Resolution in Chinese*. Ph.D. thesis, University of Pennsylvania, USA.

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

Walter Daelemans and Antal Van den Bosch. 2005. *Memory-Based Language Processing*. Cambridge University Press, Cambridge, UK.

Ido Dagan and Alon Itai. 1990. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 330–332.

Hal Daumé III and Daniel Marcu. 2005. A large-scale exploration of effective global features for a joint entity detection and tracking model. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, pages 97–104.

Arthur Dempster. 1968. A generalization of Bayesian inference. *Journal of the Royal Statistical Society*, 30:205–247.

Pascal Denis and Jason Baldridge. 2007a. Global, joint determination of anaphoricity and coreference resolution using integer programming. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 236–243.

Pascal Denis and Jason Baldridge. 2007b. A ranking approach to pronoun resolution. In *Proceedings of the Twentieth International Conference on Artificial Intelligence*, pages 1588–1593.

Pascal Denis and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 660–669.

Richard Evans. 2001. Applying machine learning toward an automatic classification of *it*. *Literary and Linguistic Computing*, 16(1):45–57.

Jenny Rose Finkel and Christopher Manning. 2008. Enforcing transitivity in coreference resolution. In *Proceedings of ACL-08: HLT, Short Papers*, pages 45–48.

Thomas Finley and Thorsten Joachims. 2005. Supervised clustering with support vector machines. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 217–224.

Yoav Freund and Robert E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296.

Nikesh Garera and David Yarowsky. 2006. Resolving and generating definite anaphora by modeling hypernymy using unlabeled corpora. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 37–44.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–170.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 44–50.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

Barbara J. Grosz. 1977. The representation and use of focus in a system for understanding dialogs. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pages 67–76.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 848–855.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1152–1161.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Stěpánek, Jiří Havelka, and Marie Mikulová. 2006. The Prague Dependency Treebank 2.0. In *Linguistic Data Consortium*.

Sanda Harabagiu, Răzvan Bunescu, and Steven Maiorano. 2001. Text and knowledge mining for coreference resolution. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 55–62.

Laura Hasler, Constantin Orasan, and Karin Naumann. 2006. NPs for events: Experiments in coreference annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1167–1172.

Peter Heeman and James Allen. 1995. The TRAINS spoken dialog corpus. CD-ROM, Linguistic Data Consortium.

Graeme Hirst. 1981. Discourse-oriented anaphora resolution in natural language understanding: A review. *American Journal of Computational Linguistics*, 7(2):85–98.

Jerry Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

Véronique Hoste and Walter Daelemans. 2005. Comparing learning approaches to coreference resolution. There is more to it than bias. In *Proceedings of the ICML Workshop on Meta-Learning*.

Véronique Hoste. 2005. *Optimization Issues in Machine Learning of Coreference Resolution*. Ph.D. thesis, University of Antewerp, Belgium.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Zhiheng Huang, Guangping Zeng, Weiqun Xu, and Asli Celikyilmaz. 2009. Accurate semantic class classifier for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1232–1240.

Nancy Ide and Dan Cristea. 2000. A hierarchical account of referential accessibility. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 416–424.

Ryu Iida, Kentaro Inui, Hiroya Takamura, and Yuji Matsumoto. 2003. Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the EACL Workshop on The Computational Treatment of Anaphora*.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2006. Exploting syntactic patterns as clues in zero-anaphora resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 625–632.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2007a. Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Transactions on Asian Language Information Processing*, 6(4).

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007b. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the ACL Workshop 'Linguistic Annotation Workshop'*, pages 132–139.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2009. Capturing salience with a trainable cache model for zero-anaphora resolution. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 647–655.

Ryu Iida. 2007. *Combining Linguistic Knowledge and Machine Learning for Anaphora Resolution*. Ph.D. thesis, Nara Institute of Science and Technology, Japan.

Heng Ji, David Westbrook, and Ralph Grishman. 2005. Using semantic relations to refine coreference decisions. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, pages 17–24.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In Bernhard Scholkopf and Alexander Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 44–56. MIT Press.

Andrew Kehler, Douglas Appelt, Lara Taylor, and Aleksandr Simma. 2004a. Competitive self-trained pronoun interpretation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 33–36.

Andrew Kehler, Douglas Appelt, Lara Taylor, and Aleksandr Simma. 2004b. The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Human Language Technologies 2004: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 289–296.

Andrew Kehler. 1997. Probabilistic coreference in information extraction. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 163–173.

Christopher Kennedy and Branimir Boguraev. 1996. Anaphor for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 113–118.

Manfred Klenner and Étienne Ailloud. 2008. Enhancing coreference clustering. In *Proceedings of the Second Workshop on Anaphora Resolution*, pages 31–40.

Manfred Klenner and Étienne Ailloud. 2009. Optimization in coreference resolution is not needed: A nearly-optimal algorithm with intensional constraints. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 442–450.

Manfred Klenner. 2007. Enforcing consistency on coreference sets. In *Proceedings of Recent Advances in Natural Language Processing*.

Fang Kong, GuoDong Zhou, and Qiaoming Zhu. 2009. Employing the centering theory in pronoun resolution from the semantic perspective. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 987–996.

Shalom Lappin and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–562.

Xiaoqiang Luo and Imed Zitouni. 2005. Multi-lingual coreference resolution with syntactic features. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, pages 660–667.

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the Bell tree. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 135–142.

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

Xiaoqiang Luo. 2007. Coreference or not: A twin model for coreference resolution. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 73–80.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Katja Markert and Malvina Nissim. 2005. Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, 31(3):367–402.

Andrew McCallum and Ben Wellner. 2003. Toward conditional models of identity uncertainty with application to proper noun coreference. In *Proceedings of the IJCAI Workshop on Information Integration on the Web*.

Andrew McCallum and Ben Wellner. 2004. Conditional models of identity uncertainty with application to noun coreference. In *Advances in Neural Information Proceesing Systems*.

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

Ruslan Mitkov. 1999. Anaphora resolution: The state of the art. Technical Report (Based on the COLING/ACL-98 tutorial on anaphora resolution), University of Wolverhampton, Wolverhampton.

Ruslan Mitkov. 2001. Outstanding issues in anaphora resolution. In Al. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 110–125. Springer.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman.

Natalia N. Modjeska, Katja Markert, and Malvina Nissim. 2003. Using the web in machine learning for other-anaphora resolution. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 176–183.

MUC-6. 1995. *Proceedings of the Sixth Message Understanding Conference*.

MUC-7. 1998. *Proceedings of the Seventh Message Understanding Conference*.

Christoph Müller, Stefan Rapp, and Michael Strube. 2002. Applying co-training to reference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 352–359.

Christoph Müller. 2006. Automatic detection of nonreferential it in spoken multi-party dialog. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 49–56.

Vincent Ng and Claire Cardie. 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.

Vincent Ng and Claire Cardie. 2002b. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 730–736.

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

Vincent Ng. 2004. Learning noun phrase anaphoricity to improve conference resolution: Issues in representation and optimization. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 151–158.

Vincent Ng. 2007a. Semantic class induction and coreference resolution. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 536–543.

Vincent Ng. 2007b. Shallow semantics for coreference resolution. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, pages 1689–1694.

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

Vincent Ng. 2009. Graph-cut-based anaphoricity determination for coreference resolution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 575–583.

Giang Linh Nguy, Václav Novák, and Zdeněk Žabokrtský. 2009. Comparison of classification and ranking approaches to pronominal anaphora resolution in Czech. In *Proceedings of the SIGDIAL 2009 Conference*, pages 276–285.

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*, pages 275–283.

Kristina Nilsson. 2010. *Hybrid Methods for Coreference Resolution in Swedish*. Ph.D. thesis, Stockholm University, Sweden.

Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. 2002. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 82–86.

Constantin Orăsan and Richard Evans. 2007. NP animacy identification for anaphora resolution. *Journal of Artificial Intelligence Research*, 29:79 – 103.

Constantin Orăsan, Dan Cristea, Ruslan Mitkov, and António H. Branco. 2008. Anaphora Resolution Exercise: An overview. In *Proceedings of the 6th Language Resources and Evaluation Conference*, pages 2801–2805.

Chris Paice and Gareth Husk. 1987. Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun 'it'. *Computer Speech and Language*, 2:109–132.

Massimo Poesio and Mijail A. Kabadjov. 2004. A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 663–668.

Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004a. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 143–150.

Massimo Poesio, Olga Uryupina, Renata Vieira, Mijail Alexandrov-Kabadjov, and Rodrigo Goulart. 2004b. Discourse-new detectors for definite description resolution: A survey and a preliminary proposal. In *Proeedings of the ACL Workshop on Reference Resolution*.

Massimo Poesio, David Day, Ron Artstein, Jason Duncan, Vladimir Eidelman, Claudio Giuliano, Rob Hall, Janet Hitzeman, Alan Jern, Mijail Kabadjov, Stanley Yong Wai Keong, Gideon Mann, Alessandro Moschitti, Simone Ponzetto, Jason Smith, Josef Steinberger, Michael Strube, Jian Su, Yannick Versley, Xiaofeng Yang, and Michael Wick. 2007. ELERFED: Final report of the research group on Exploiting Lexical and Encyclopedic Resources For Entity Disambiguation. Technical report, Summer Workshop on Language Engineering, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD.

Simone Paolo Ponzetto and Massimo Poesio. 2009. State-of-the-art NLP approaches to coreference resolution: Theory and practical recipes. In *Tutorial Abstracts of ACL-IJCNLP 2009*, page 6.

Simone Paolo Ponzetto and Michael Strube. 2006a. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Human Language Technologies 2006: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 192–199.

Simone Paolo Ponzetto and Michael Strube. 2006b. Semantic role labeling for coreference resolution. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 143–146.

Simone Paolo Ponzetto and Michael Strube. 2007. Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30:181–212.

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*, pages 650–659.

Andrei Popescu-Belis, Loïs Rigouste, Susanne Salmon-Alt, and Laurent Romary. 2004. Online evaluation of coreference resolution. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1507–1510.

Long Qiu, Min-Yen Kan, and Tat-Seng Chua. 2004. A public reference implementation of the RAP anaphora resolution algorithm. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 291–294.

John Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

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

Marta Recasens and M. Antónia Martí. 2009. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 43(4).

Marta Recasens, Toni Martí, Mariona Taulé, Lluís Màrquez, and Emili Sapena. 2009. SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions* (SEW-2009), pages 70–75.

Candace Sidner. 1979. *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*. Ph.D. thesis, Massachusetts Institute of Technology, USA.

Wee Meng Soon, Hwee Tou Ng, and Chung Yong Lim. 1999. Corpus-based learning for noun phrase coreference resolution. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 285–291.

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.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 656–664.

Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom. 2010. Coreference resolution with Reconcile. In *Proceedings of the ACL 2010 Conference Short Papers*.

Michael Strube, Stefan Rapp, and Christoph Müller. 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.

Michael Strube. 2002. NLP approaches to reference resolution. In *Tutorial Abstracts of ACL 2002*, page 124.

Michael Strube. 2009. Anaphernresolution. In *Computerlinguistik und Sprachtechnologie. Eine Einfuhrung*. Springer, Heidelberg, Germany, 3rd edition.

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The tüba-d/z treebank: Annotating German with a context-free backbone. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 2229–2235.

Joel Tetreault. 2005. *Empirical Evaluations of Pronoun Resolution*. Ph.D. thesis, University of Rochester, USA.

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

Olga Uryupina. 2004. Linguistically motivated sample selection for coreference resolution. In *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium.*

Kees van Deemter and Rodger Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637.

Yannick Versley, Alessandro Moschitti, Massimo Poesio, and Xiaofeng Yang. 2008a. Coreference systems based on kernels methods. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 961–968.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008b. BART: A modular toolkit for coreference resolution. In *Proceedings of the ACL-08: HLT Demo Session*, pages 9–12.

Yannick Versley. 2006. A constraint-based approach to noun phrase coreference resolution in German newspaper text. In *Konferenz zur Verarbeitung Natürlicher Sprache*.

Yannick Versley. 2007. Antecedent selection techniques for high-recall coreference resolution. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 496–505.

Renata Vieira and Massimo Poesio. 2000. Processing definite descriptions in corpora. In S. Botley and A. McEnery, editors, *Corpus-based and Computational Approaches to Discourse Anaphora*, pages 189–212. UCL Press.

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

Marilyn Walker, Aravind Joshi, and Ellen Prince, editors. 1998. *Centering Theory in Discourse*. Oxford University Press.

Holger Wunsch. 2010. *Rule-based and Memory-based Pronoun Resolution for German: A Comparison and Assessment of Data Sources*. Ph.D. thesis, University of Tübingen, Germany.

Xiaofeng Yang and Jian Su. 2007. Coreference resolution using semantic relatedness information from automatically discovered patterns. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 528–535.

Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2003. Coreference resolution using competitive learning approach. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 176–183.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2004a. Improving noun phrase coreference resolution by matching strings. In *Proceedings of the First International Joint Conference on Natural Language Processing*, pages 22–31.

Xiaofeng Yang, Jian Su, GuoDong Zhou, and Chew Lim Tan. 2004b. An NP-cluster based approach to coreference resolution. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 226–232.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2005. Improving pronoun resolution using statistics-based semantic compatibility information. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 165–172.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2006. Kernel based pronoun resolution with structured syntactic knowledge. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 41–48.

Xiaofeng Yang, Jian Su, Jun Lang, Chew Lim Tan, and Sheng Li. 2008a. An entity-mention model for coreference resolution with inductive logic programming. In *Proceedings of ACL-08: HLT*, pages 843–851.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2008b. A twin-candidate model for learning-based anaphora resolution. *Computational Linguistics*, 34(3):327–356.

Dmitry Zelenko, Chinatsu Aone, and Jason Tibbetts. 2004. Coreference resolution for information extraction. In *Proceedings of the ACL Workshop on Reference Resolution and its Applications*, pages 9–16.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and resolution of Chinese zero pronouns: A machine learning approach. In *Proceedings of the 2007 Joint Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning*, pages 541–550.

GuoDong Zhou and Fang Kong. 2009. Global learning of noun phrase anaphoricity in coreference resolution via label propagation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 978–986.