Published in final edited form as:
J Biomed Inform. 2011 December ; 44(6): 1113–1122. doi:10.1016/j.jbi.2011.08.006.

Coreference resolution: A review of general methodologies and applications in the clinical domain

Jiaping Zheng[a,*], **Wendy W. Chapman**[b], **Rebecca S. Crowley**[c], and **Guergana K. Savova**[a,d]

[a]Children's Hospital Boston, 300 Longwood Ave, Boston, MA 02115, United States

[b]University of California, San Diego, 9500 Gilman Dr., Bldg 2 #0728, La Jolla, CA 92093, United States

[c]University of Pittsburgh Medical Center, 5150 Centre Ave, Pittsburgh, PA 15232, United States

[d]Harvard Medical School, 25 Shattuck Street, Boston, MA 02115, United States

Abstract

Coreference resolution is the task of determining linguistic expressions that refer to the same real-world entity in natural language. Research on coreference resolution in the general English domain dates back to 1960s and 1970s. However, research on coreference resolution in the clinical free text has not seen major development. The recent US government initiatives that promote the use of electronic health records (EHRs) provide opportunities to mine patient notes as more and more health care institutions adopt EHR. Our goal was to review recent advances in general purpose coreference resolution to lay the foundation for methodologies in the clinical domain, facilitated by the availability of a shared lexical resource of gold standard coreference annotations, the Ontology Development and Information Extraction (ODIE) corpus.

Keywords

NLP; Coreference; EHR

1. Introduction

Coreference resolution is the task of determining linguistic expressions that refer to the same real-world entity in natural language. For example, in the sentences "Have reviewed the electrocardiogram. It shows a wide QRS with a normal rhythm but no delta waves." the phrases "the electrocardiogram" and "It" refer to the same entity, i.e. the electrocardiogram.

It has been widely acknowledged that the unstructured clinical narratives are a rich source of information that complements the structured data in the electronic health record (EHR). Applying natural language processing (NLP) technologies to extract information from the narratives can not only unlock information that is only present in the free text portion of the EHR but also improve performance when combined with structured data. Fiszman et al. [1] extracted clinical information from ventilation/perfusion lung scan reports, which is only available in free text format. Xu et al. [2] devised a medication extraction system that achieved over 90% F-measure on drug names and signatures, which are otherwise absent in coded data. Zeng et al. [3] found that combining an NLP system and the ICD-9 codes

[*]Corresponding author. jiaping.zheng@childrens.harvard.edu (J. Zheng), wendy.w.chapman@gmail.com (W.W. Chapman), crowleyrs@upmc.edu (R.S. Crowley), guergana.savova@childrens.harvard.edu (G.K. Savova)..

improves accuracy, sensitivity and specificity in a study to extract principal diagnosis from discharge summaries. Li et al. [4] concluded that NLP systems provide information that is not present in the structured data. Liao et al. [5] achieved higher positive predictive value in defining a rheumatoid arthritis cohort by utilizing the clinical narrative data. Kullo et al. [6] leveraged the unstructured information in the EHR (smoking status and medication dosage, frequency, and route) to conduct genome-wide association study of peripheral arterial disease (PAD). Savova et al. [7] demonstrated the utility of NLP in classifying PAD status.

However, to take full advantage of the information in the clinical free text, coreference resolution is an indispensable component. Coreference serves the critical role of linking related information together. Garla et al. [8] identified that lack of coreference resolution contributed to misclassifications in a clinical document classification system. Consider the short snippet in Example 1 (Table 1) from a clinical note. Without a coreference algorithm to establish that “significant pain in the shoulder” and “his discomfort” corefer, one would not be able to conclude that the patient uses Tylenol to treat his shoulder pain.

Attributes, temporal descriptions, and contextual information necessary for understanding whether conditions, symptoms, and treatments have occurred or are merely planned are often spread over several sentences or even paragraphs rather than within a single sentence and require coreference resolution for accurate interpretation.

For example, accurate assignment of attributes to named entities (Examples 2 and 5), accurate assignment of temporal information to an event (Example 3), distinguishing planned events from events that occurred (Example 4) can only be achieved by resolving the coreferential phrases.

In Example 2, a system needs to separate the adenoma ($\{m_2,m_4\}$) and the carcinoma ($\{m_1,m_3,m_5\}$) through coreference resolution to collect all the attributes of each of the entities.

In Example 3, the quality that the chest discomfort occurs at rest and lasts 30 min requires the resolution of the two highlighted phrases.

By relating “that variety” to “back extensor strengthening exercises” in Example 4, a system can determine that the physician is planning a home program of the back extensor strengthening exercises.

Resolving the three highlighted phrases in Example 5 is critical as they are the force that holds the other pieces of information together. Only after linking the three phrases can one ascertain that symptoms of nausea and vomiting have occurred earlier but wors-ened recently.

Armed with a textual coreference resolution system, a higher-level system can resolve coreference between the narrative notes and the structured data to yield a richer picture. For example, such a system can link the detailed prescription and laboratory data from the EHR with the textual mentions in a clinical note. In Example 6, the start date of the Tramadol and previous dosing information can be retrieved from the structured data. But this is beyond the scope of our review. We aim at methods for coreference resolution in text.

Coreference resolution has long been recognized as a difficult task. Research in the general English domain dates back to 1960s and 1970s [9, chapter 3]. Various systems from heuristics-based ones to statistical ones have been developed. In particular, there have been growing efforts since the 6th and 7th Message Understanding Conferences (MUC) [10,11] and the Automatic Content Extraction (ACE) program[1] initiated shared tasks on coreference resolution and released their annotated corpora in the last two decades. However, the clinical domain has not seen major development, which can be partially attributed to the lack of sharable annotated clinical text. The recent US government initiatives that promote the use of electronic health records provide opportunities to mine patient notes as more and more health care institutions adopt EHR. In this paper we give a review of the approaches in the general English and biomedical literature domains and discuss challenges in applying those techniques in the clinical narrative.

1.1. Related work

Hirst [9] provided a survey of research on anaphora during the early years. The approaches, mostly heuristic-based, have largely been superseded since the 1990s. Trends of transition of research focus from heuristics to statistical and machine learning approaches can be seen in Mitkov [12]. Ng [13] concentrated exclusively on supervised machine learning approaches that started in the mid-1990s. Our survey is not limited to particular methodologies, and has a focus on clinical applications.

1.2. Definitions

Formally, coreference consists of two linguistic expressions—antecedent and anaphor. The *anaphor* is the expression whose interpretation (i.e., associating it with an either concrete or abstract real-world entity) depends on that of the other expression. The *antecedent* is the linguistic expression on which an anaphor depends. In the first example in Section 1, “the electrocardiogram” is the antecedent, and “It” is the anaphor. Similarly in Example 1, “significant pain in the shoulder” is the antecedent, and “his discomfort” is the anaphor. The relationship between the antecedent and the anaphor is usually “identity”—they both refer to the same entity. A broader concept of *anaphora* includes a pair of linguistic expressions whose relationship does not have to be identity.

These linguistic expressions, the antecedents and the anaphors, are collectively called *markables* in the MUC corpus. Two coreferring markables form a *pair*, while one or more pairs that refer to the same entity form a *chain*. In the ACE corpus, the linguistic expressions are called *mentions*, and the entities these mentions refer to are, naturally, *entities*.

The coreference resolution task is to discover the antecedent for each anaphor in a document. Since the coreference relation is transitive, the set of all the transitive closures of the markables forms a partition, in other words, a set that contains the sets of markables in each chain. For text processing systems, such as information retrieval (IR) and information extraction (IE), identifying the exact antecedent is less important than correctly partitioning the markables. For instance, to extract all the relevant information about the arthrogram in Example 7, it would be sufficient to link the second “they” to any of the other three markables, as long as the four markables are in the same set. Moreover, it is not always clear which is *the* antecedent. Therefore, most systems strive to generate a correct partition.

1.3. Coreference and the clinical narrative

The types of markables that a coreference resolution system resolve are unique to the domains. The general English domain focuses on person, location, and organization [11]. The shared task[2] in the biomedical literature domain focused on finding coreferential mentions of genes and proteins. In the clinical narrative, however, the types are mainly disorders, signs or symptoms, anatomical sites, medications, and procedures.

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

[2] https://sites.google.com/site/bionlpst/home/protein-gene-coreference-task.

In addition to the difference in the markable types, Coden et al. [14] showed that the language in the clinical notes differs from the general English. The average sentence length in clinical notes is only approximately half of that in the general English texts. The vocabulary size of clinical notes is also smaller than the general English texts. Meystre et al. [15] contrasted the clinical texts with the biomedical texts, and argued that the characteristics of clinical texts pose a special challenge to NLP. Methods for coreference resolution need to account for these subdomain language characteristics, such as the word and sentence distance between coreferential mentions. Furthermore, different genres of clinical texts show different patterns. For example, anatomical site concepts are more prevalent in procedure notes, including radiology, pathology, and operation notes, than discharge summaries.

During the past two decades, several systems have been developed to extract named entities (NEs) from clinical narrative, first specialized in certain report types [16-19], and later more general purpose [20,3,21,22]. The community is now moving towards semantic analysis and discourse processing, including relation discovery and semantic role labeling. However, there have been only a handful of efforts researching coreference in the clinical narrative.

Hahn et al. [23] included a nominal anaphora resolution algorithm in a knowledge mining system from findings reports. As part of the Ontology Development and Information Extraction project (ODIE),[3] a corpus of 100,000 words of clinical text was doubly annotated and adjudicated [24] to include 7214 markables, 5992 pairs and 1304 chains. The corpus will be made available to the research community under IRB and Data Use Agreements. As part of their work on developing a tool for cancer characteristics information extraction, Coden et al. [25] manually annotated 302 Mayo Clinic pathology notes. The annotation schema included coreference annotations for anatomical sites and histologies mapped to the International Classification of Diseases for Oncology (ICD-O) [26]. Two mentions that are exact strings and map to the same concept were annotated as coreferential. In addition, each anatomical site or histology mention is coreferenced with any instance of its parent anatomical site as defined by ICD-O. Roberts et al. [27] described their work on creating a multi-layered, semantically annotated corpus, the Clinical E-Science Framework (CLEF), in which one of the annotated relations is coreference.

Our goal was to review recent advances in general purpose coreference resolution to lay the foundation for methodologies in the clinical domain, facilitated by the availability of a shared lexical resource of gold standard coreference annotations, the ODIE corpus.

2. Material and methods

We selected publications from the Association for Computational Linguistics (ACL) Anthology[4] by querying for “anaphora” and “coreference,” but excluded papers that did not focus on the English language. The search returned about 200 results. We also selected publications using the same keywords in PubMed, but excluded papers that focused on neuroscientific or psycholinguistic discoveries. This query yielded fewer than 10 papers. Finally, publications frequently referenced in the papers from the above two sets were also included.

[3] https://bmir-gforge.stanford.edu/gf/project/odie.

[4] http://aclweb.org/anthology-new/.

3. Heuristics-based approaches

Early attempts at the coreference resolution task mainly involved heuristic approaches, motivated by linguistic theories. The general theme was to incorporate a knowledge source to prune unlikely antecedent candidates until a small set is obtained, and then select the best candidate based on the current focus [28] of attention or the preferred center. These approaches tended to employ a multitude of features, including syntactic (the gender of the two mentions must agree), semantic (a mention with the same semantic role as the anaphor is given preference), and pragmatic (the topic under discussion usually remains unchanged unless there are indications otherwise[5]) constraints and preferences. Many of them also resolved different types of anaphoric phrases at once, even some not exactly coreferential.[6] Rich and LuperFoy [29] reported on a pronominal anaphora resolution system consisting of a set of modules, each of which handles one aspect of anaphora theory. Hobbs [30] employed a deepest-first tree search procedure on the syntactic parse tree of a sentence to find the first candidate that satisfies a set of hand-crafted constraints. The search started from the immediate dominating noun phrase (NP) of the pronoun. The candidate NP antecedent was selected based on two criteria. Criterion one selected as antecedent the NP on a branch to the left of the pronoun-dominating NP path. Criterion two stated that there should be another NP between the candidate from criterion one and the dominating NP. Both criteria one and two had to be satisfied. If no matching candidate was found, the algorithm traversed up the tree, and broadened the search. The accuracy on a small test set was between 88.3% and 91.7%. Lappin and Leass [31] used a heuristic approach to resolve pronouns and lexical anaphors (reflexives and reciprocals). The Resolution of Anaphora Procedure (RAP) algorithm, as it is referred to, operates on salience measures derived from syntactic structure and an attentional state model. They achieved 86% accuracy. Whereas RAP requires a full syntactic parser, Kennedy and Boguraev [32] presented an extension to it that substitutes the parse tree with part of speech, phrasal, and other morphosyntactic features. The accuracy of their system was 75.4%. Castaño et al. [33] described a system to resolve pronominal phrases and bio-type noun phrases in the biomedical literature. In their system, a potential antecedent is assigned a salience measure based on a series of criteria, indicating the "compatibility" of the anaphor and the antecedent. The one(s) with the highest salience measure is selected as antecedent(s). The precision and recall are 77% and 72% respectively.

4. Supervised approaches

In the mid-1990s, methods for performing supervised coreference resolution sprang up. The widespread availability of the MUC and ACE corpora further shaped the research community to move towards statistical approaches. Complete heuristics-based systems gradually saw a decline of interest in the community, although isolated rules are still employed to encode hard linguistic constraints. Two types of machine learning methods emerged—a two-step binary classification followed by clustering and a ranking approach. The key distinction between them is that the binary classification approach makes coreference decisions on the antecedent candidates independently of each other, while the ranking approach takes into account other antecedent candidates.

[5] For example, in "When Sue went to Nadia's home for dinner, *she* ate sukiyaki au gratin." we know "she" refers to Sue, not Nadia, because Sue is the topic in the preceding clause, and remains unchanged as there is no other construction that introduces a new topic [9].

[6] For example, the contrastive use of one-anaphora in "a big green pyramid and a small *one*"[9].

4.1. Binary classification

The binary classification approach involves two steps. First, for a given anaphor, the classifier determines for each candidate antecedent whether the anaphor corefers with the antecedent. A clustering algorithm then takes these pairwise coreference decisions and generates a partition of the set of all markables in the document, such that all the markables in each partition refer to the same entity. This process is named the “mention-pair” model, since it hinges on a pair of markables (mentions).

A different but similar approach is the “entity-mention” model. It also casts the task as a binary classification problem, except that the classifier predicts whether a markable is coreferent with a partially-formed entity (chain), instead of a single markable as in the “mention-pair” model. The second clustering step proceeds in an analogous manner.

4.1.1. Mention-pair model—McCarthy and Lehnert [34] were among the first to adopt a machine learning approach to resolving coreference. They evaluated a decision-tree-based system on the MUC-5 English Joint Venture corpus. The system was trained on all possible pairs in the training set, with eight features.[7] The result outperformed an earlier heuristics-based system. Numerous systems were subsequently developed, and generally followed this paradigm.

One of the limitations acknowledged by the authors regarding their study is that the imbalance of the positive and negative training instances causes a bias towards classifying more negative pairs. Because all possible pairings of markables are extracted, the negative instances far outnumbered the positive ones. An influential method to creating training instances to mitigate this problem was proposed in Soon et al. [35]. Positive instances were created from a markable and its immediate preceding markable that are coreferent. For every positive instance that involves markables m_i and m_j, negative instances were created for each pair of markables m_k and m_j, where $i < k < j$. A variant of this method that differs slightly in the creation of positive instances was proposed in Ng and Cardie [36], whereby the immediate preceding non-pronominal markable is paired with a non-pronominal markable to create a positive instance. Another variant in the creation of negative instances was described in Ng and Cardie [37]. For every anaphoric markable m_j whose farthest antecedent markable to the left is m_i, a negative instance was created for each markable m_k such that $i < k < j$ and m_k and m_j are not coreferent.

Other methods in reducing the training instances focused on removing obvious negative instances to improve the training set balance or removing elusive positive instances to help the algorithm to learn from “confident” pairs. Yang et al. [38] removed markables that violate gender, number or person agreement with the anaphor. Harabagiu et al. [39] crafted rules manually to remove hard positive instances (such as those that require external knowledge) while preserving the coverage of chains (based on the transitivity nature of the coreference relation) as much as possible. Ng and Cardie [37] used a learner to exclude hard positive instances. Uryupina [40] employed different methods in eliminating irrelevant or hard positive instances for pronoun, proper name, definite NP, and other types of anaphoric markables.

The number of features obtained from the training instances varies considerably, from a small set of eight [34] to nearly 40 [36]. Uryupina [40] even reported 187 features. The features can either operate on one of the two markables or both of them. Most of these

[7] These features include whether each markable contains a name, refers to a joint venture child, whether one markable contains a reference to the other, whether both markables refer to a joint venture child, whether the two markables share a common noun phrase, and whether they are in the same sentence.

features fall into one of the categories of lexical, syntactic, or semantic. Common features are summarized in Table 2.

Lexical features mainly include string matching operations, such as exact match, substring match, and overlapping words. Syntactic features consist of grammatical roles, phrasal types, linguistic constraints like agreement and binding theory. Most of these syntactic features are derived from the parse trees in a heuristic manner. A notable exception is that Yang et al. [46] utilized the parse trees directly as a structured feature. Semantic features usually involve consulting an external ontology, for example WordNet [47]. Ng [48] experimented with sophisticated semantic features but found limited performance gains, due to the difficulty in accurately computing these features. Bengtson and Roth [49] evaluated the contributions of the features commonly used.

Decision tree [34-36], maximum entropy/logistic regression [50-52,43], support vector machine [53], generative statistical model [54], averaged perceptron [49] and conditional random fields [55] have all been reported in the literature. Justification for decision trees is usually its ease of interpretation for humans [56], while the reason for the choice of maximum entropy is that they are able to handle potentially non-independent features [50].

4.1.2. Entity-mention model—A common critique of the mention-pair model is that it cannot capture information beyond the mention pair. Consider a pair of a non-pronominal antecedent and a pronominal anaphor. The information that can be obtained from the two markables to determine their coreferential status is very limited, except for the gender and number agreements. Discarding these hard-to-resolve instances as discussed earlier may help the algorithms to learn from other strong evidence. However, this type of pair is a very frequent linguistic phenomenon.

In light of this shortcoming of the mention-pair model, Yang et al. [57] presented an approach to determining whether a noun phrase is coreferential with an existing (partial) coreferential cluster. They obtained better results on the GENIA data set [58] than the mention-pair model using a decision tree system.

Training instances in an entity-mention model encompasses an anaphor and a cluster of preceding NPs. Instances are created similarly to the mention-pair model, i.e., for each positive instance, negative instances are created with the anaphora and its non-coreferential clusters.

In addition to features used in the mention-pair models describing the relationships between the anaphor and its antecedent, features encoding relationships between an anaphor and a partial cluster are added. Table 3 lists features proposed in Culotta et al. [52]. These cluster-level features utilize first-order logic to expand upon the pairwise features. For example, the number agreement feature (whether the two markables are both singular, or plural, or one is singular and the other plural) between the antecedent and anaphor in the mention-pair model can be transformed to the number agreement among the anaphor and *all* [57] or *any* [59,41] of the NPs in the cluster.

Similar sets of classification methods are employed with these features, including decision tree [57] and maximum entropy [41]. One unique method is proposed in Yang et al. [59]—inductive logic programming. Training instances are represented as predicates. For example, a predicate $link(e_{i\_j},m_j)$ encodes that mention m_j is coreferential with partial entity i before the jth mention. A feature that indicates the number agreement can be represented as $entNum\text{-}Agree(e_{i\_j},m_j,v)$, where v is an indicator variable. The system takes these predicates and induces a set of rules to classify new instances.

4.1.3. Partitioning—The binary classification results from the mention-pair model or the entity-mention model are only the first step in resolving coreference. The markables need to be clustered into chains based on the predictions from the classifier.

For example, for the markables in Example 8, a system may generate the following results for the pairs: $\langle m_1,m_2\rangle$ non-coreferential, $\langle m_1,m_3\rangle$ non-coreferential, $\langle m_2,m_3\rangle$ non-coreferential, $\langle m_2,m_4\rangle$ non-coreferential, $\langle m_2,m_5\rangle$ coreferential, etc. The partitioning algorithm is responsible to cluster the five markables into three sets: $\{m_1\}$, $\{m_2,m_4\}$, and $\{m_3,m_5\}$, from the imperfect classification results.

Let $\langle m_i,m_j\rangle^+$ denote that the classification output for markables m_i and m_j is coreferential, and $\langle m_i,\mathrm{m_j}\rangle^-$ otherwise. Suppose the results for four markables $m_1 \ldots m_4$ are $\langle m_1,m_2\rangle^+$, $\langle m_1,m_3\rangle^-$, $\langle m_1,m_4\rangle^+$, $\langle m_2,m_3\rangle^+$, $\langle m_2,m_4\rangle^-$, and $\langle m_3,m_4\rangle^+$. There is not a natural grouping of the four markables that is consistent with the pairwise result.

Similarly in the entity-mention case, let $\langle\{m_i \ldots m_j\}, m_k\rangle^\pm$ denote the classification result for a partial cluster $\{m_i \ldots m_j\}$ and a markable m_k. Given results of $\langle\{m_1\},m_2\rangle^-$, $\langle\{m_2\},m_3\rangle^+$, and $\langle\{m_1,m_3\},m_4\rangle^+$, there is not a partition of the four markables that satisfies the three individual results.

Two greedy algorithms (closest-first and best-first) are commonly used. The closest-first algorithm links (in the mention-pair model) the closest antecedent that the classifier predicts positive to the anaphor [35]. If all instances for which a markable is tested as antecedents are negative, this markable is considered non-anaphoric. In a classifier that generates a probability for its predictions, the best-first algorithm selects the candidate antecedent with the highest probability (usually over a threshold δ of 0.5) as the final choice [36]. If all instances for a markable are below the threshold, the markable is considered non-anaphoric. Ng and Cardie [36] showed that the best-first algorithm gives better results than the closest-first algorithm. In the entity-mention model, the algorithms select the closest or the best partial cluster to link an anaphor [57].

The simple greedy clustering algorithms only make use of a subset of the classification results to link markables to its antecedent or partial cluster. This approach leads to a bias towards the positive classification results. Using the earlier example of mention-pair model in this section ($\langle m_1,m_2\rangle^+$, $\langle m_1,m_3\rangle^-$, $\langle m_1,m_4\rangle^+$, $\langle m_2,m_3\rangle$, $\langle m_2,m_4\rangle^-$, and $\langle m_3,m_4\rangle^+$), the cluster algorithm link m_2–m_1 and m_3–m_2. Based on the transitivity property of coreference relations, it follows that m_1 and m_3 are coreferential, which contradicts with the classification result for this pair. This would happen no matter how unlikely it is for this pair to be coreferential according to the classification result.

Globally optimized clustering algorithms have been proposed to address this problem. Denis and Baldridge [45] used integer linear programming to find a linking scheme that maximally agrees with the classification results, both positive and negative. This is achieved by minimizing an objective function

$$\sum_{\langle i,j\rangle\in M\times M} -\log p\cdot\chi_{\langle i,j\rangle} - \log\left(1-p\right)\cdot\left(1-\chi_{\langle i,j\rangle}\right) \tag{1}$$

subject to

$$\chi_{\langle i,j\rangle}\in\{0,1\} \qquad \forall\,\langle i,j\rangle\in M\times M \tag{2}$$

where M is the set of all markables, and p is the probability given by the coreference classifier that m_i and m_j corefer.

Luo et al. [41] cast the clustering step as a search problem in a search space represented by a Bell tree—the ith level of the tree corresponds to all the possible partitions of the first i markables in the document. Thus, the leaf nodes represent all possible complete partitions of the markables. The algorithm searches for a path from the root to a leaf node that optimizes a maximum entropy model from the mention-pair model or the entity-mention model. Since the number of leaf nodes (given by the Bell number $B_n = \frac{1}{e}\sum_{k=0}^{\infty}\frac{k^n}{k!}$ for n markables) grows rapidly as n increases, poorly-scored children and nodes that violate certain constraints are pruned.

Nicolae and Nicolae [60] represented the clustering problem in an undirected graph. The nodes in the graph represent the markables, and the weights on the edges are derived from the classification probabilities. They designed a graph partitioning algorithm to find the best clustering of the markables from the graph.

4.2. Ranking and beyond

A drawback of both the mention-pair model and the entity-mention model is that they consider the candidates (markables in the mention-pair model, partial chains in the entity-mention model) independently. They cannot measure how likely a markable is the antecedent for a given anaphor, relative to the other candidate markables. Ranking models are designed to address this issue.

A precursor to ranking models is described in Yang et al. [38], where an instance is created from an anaphor and two candidates, one of which is the true antecedent, and the other is not. In this twin-candidate model, markables are compared in a pairwise fashion. The best overall markable is the one that wins the most round robin competitions.

Contrary to the twin-candidate model, Denis and Baldrdige [43] considered all candidates at once in a log-linear model:

$$P(\alpha_i|\pi) = \frac{1}{Z}\exp\sum_{j=1}^{m} w_j f_j(\pi, \alpha_i)$$

where π is the anaphor, α_i is an antecedent candidate, $f_j(\pi,\alpha_i)$ is a feature computed from π and α_i, w_j is the weight, and Z is a normalization factor. The candidate with the highest probability is taken as the final antecedent. A preferable property of this method is that it essentially obviates the need for a clustering algorithm, as it innately captures the competition among the candidates.

Analogous to entity-mention model's improvement upon the mention-pair model by incorporating information from other mentions in a cluster, Rahmand and Ng [61] introduced a cluster ranking to improve the performance of the mention ranking model by taking advantage of information in a cluster.

There are a few other methods that do not fall in any of the above categories. Finelye and Joachims [62] learned a similarity metric between pairs of markables, and applied correlation clustering [63] to maximize the sum of the similarity scores for markables in the same cluster (chain). McCallum and Wellner [55] also eliminated the classification step by treating the task as a graph partitioning problem (using correlation clustering), where the vertices are the markables, and the undirected edges' weights are the clique potentials on the two vertices. Daumé III and Marcu [64] modeled the named entity recognition and coreference resolution simultaneously in an online learning model using the Learning as Search Optimization framework [65]. Culotta et al. [52] further expanded the entity-mention model by determining coreferential status between two clusters, exploiting complex first-order logic features. Ng [42] trained a support vector machine ranker of partitions generated by 54 different systems.

As a summary, Table 4 gives an overview of the various methods reviewed in this section, and their performance. However, it should be noted that system evaluations are performed on different corpora, and results are reported in different metrics. Therefore, the actual figures are not directly comparable. Section 6 provides a more in-depth analysis of the issues associated with evaluation.

4.3. Anaphoricity

There is another thread of research that focuses on distinguishing between anaphoric and non-anaphoric phrases, commonly referred to as "anaphoricity". Rather than relying on the clustering algorithm to use the classification results implicitly to identify non-anaphoric markables, an anaphoricity classifier serves as a filter to the classifier—only markables that are determined to be anaphoric by the anaphoricity classifier are included to create test instances and passed to the classification algorithm.

Although not essential for the binary-classification-based methods, the anaphoricity classifier is critical in a ranking model, because the ranker categorically links an anaphor to one of its candidate antecedents. Therefore, ranking models are commonly combined with an anaphoricity classifier [43,61].

Previously, efforts are limited to ruling out expletive "it"[8] with both heuristic rules [31] and machine learning [67]. This additional layer of processing is later extended to include other non-anaphoric NPs. Bean and Riloff [68] built a system with four types of heuristic rules. Ng and Cardie [56] trained a decision tree classifier to determine NP anaphoricity, and limited the coreference classifier to only consider NPs that are classified as anaphoric by the anaphoricity classifier. The result demonstrated a significant improvement in precision. However, this improvement was offset by the drop in recall, which led to poorer F-score. Additional heuristics-based constraints reduced the drop in recall, and reverted the trend in decreasing Fscore. Ng [69] explored other possible configurations of the interactions between the anaphoricity and coreference classifiers, namely, incorporating the anaphoricity result as a feature in the coreference classifier, and optimizing the overall performance instead of two separate classifiers. Their experiment showed significant gains in F-score when using the anaphoricity classifier as a filter and optimizing globally.

A more global optimization of the anaphoricity classifier is presented in Denis and Baldridge [45]. The method is an extension to the integer linear programming algorithm used by the same authors to cluster markables into chains. The objective function (Eq. 1) is augmented with requirements to resolve and only resolve anaphors, and constraints (Eq. 2) are expanded to enforce consistency among the results.

[8] Expletives are words that fulfill syntactic requirements but do not carry meanings. For example, "*It* is important that the patient receive a follow up exam."

4.4. Specialized models

It is worth noting that many coreference resolution applications focus on a single type of NP, or handle different types of NPs separately, based on the observation that different types of NPs exhibit different patterns in terms of coreference participation [70,71]. Strube et al. [72] provided empirical evidence by examining the performance of the same set of features on different types of NPs, and obtained disparate results on pronouns, proper names, and definite NPs (in the order from highest to lowest). Ge et al. [54], and Yang et al. [46] built systems that only resolve pronouns. Morton [51] trained a maximum entropy model to resolve pronouns, and applied simple string matching to resolve proper nouns. Denis and Baldridge [43] learned separate models for third person pronouns, speech pronouns (first and second person), proper names, definite NPs, and other anaphoras that do not fall into one of the previous categories. Zelenko et al. [73] also trained five classifiers to handle names, nominal NPs, first person pronouns, second person pronouns, “it”, singular third person pronouns, and plural third person pronouns. Bergsma et al. [74] learned non-referential “it” by examining how likely “it” can be substituted with other words.

5. Unsupervised approaches

Unsupervised approaches to coreference resolution are a more recent development, and systems are still rare, although there are reports as early as 1999 [75]. However, the results are less satisfactory.

The first substantial effort to tackle the coreference resolution task in an unsupervised manner is described in Haghighi and Klein [76]. They adopted a fully generative, nonparametric Bayesian model, based on hierarchical Dirichlet processes. For each document, the goal was to find the assignment of the entity indices $\boldsymbol{Z}$ for all the mentions $\boldsymbol{X}$ that maximizes the posterior probability $P(\boldsymbol{Z}|\boldsymbol{X})$. Documents are represented as mixture models, with infinite number of components, which correspond to the number of entities. An entity is drawn from a nonparametric Dirichlet process, and then the head of the mention is generated from a symmetric Dirichlet distribution. Furthermore, a pronoun head model and a salience model are designed to improve performance on pronouns by modeling additional grammatical and semantic features (gender, number, and semantic type) and recency. They achieved F-scores ranging from 62.3% to 70.3%.

Ng [77] presented a generative unsupervised model that views coreference as an Expectation-Maximization (EM) clustering process. The model operates at the document level to induce a partition (a valid clustering) of the mentions. A document D is represented by its (ordered) mention pairs, which are assumed to be generated conditionally independently of each other given the coreferential status C_{ij} between the pair of mentions m_{ij}:

$$P(D|C) = \prod_{m_{ij} \in D} P\left(m_{ij}|C_{ij}\right).$$

The pair m_{ij} is further decomposed to three groups of mutually independent features, m_{ij}^1, m_{ij}^2, and m_{ij}^3:

$$P\left(m_{ij}|C_{ij}\right) = P\left(m_{ij}^1|C_{ij}\right) P\left(m_{ij}^2|C_{ij}\right) P\left(m_{ij}^3|C_{ij}\right)$$

EM is used to iteratively estimate the model parameters Θ, which consist of P(m^1|c), P(m^2+c), and P(m^3|c):

E-step

Compute the posterior probabilities $P(C|D,\Theta)$ based on the current Θ.

M-step

Using $P(\mathrm{C}|D,\Theta)$ obtained in the E-step, find Θ' that maximizes the expected log likelihood $\sum_C P(C|D, \Theta) \log P(D,C|\Theta')$

The performance in F-score ranged from 51.6% to 62.8% using the MUC-6 metric, and 52.8% to 56.7% using the CEAF score. The details of the scoring schemes are discussed in Section 6.

Poon and Domingos [78] modeled coreference in Markov logic network (MLN) [79], which is a first-order knowledge base with a weight attached to each clause. At its basis, the MLN is similar to Haghighi and Klein [76] in that it utilizes a head mixture model. It includes a mixture component prior, represented by the clause

```
InClust(+m,+c),
```

and the head distribution is represented by

```
InClust(m,+c) ∧ Head(m,+t).
```

The predicate

```
InClust(+m, + c)
```

is true iff mention *m* is in cluster *c*;

```
Head(m, t)
```

is true iff token *t* is the head of mention *m*. The '+' sign signifies that the MLN contains an instance of the rule, with a separate weight, for each value combination of the variables with a plus sign. Additional predicates are introduced to address pronouns, apposition and predicate nominals. Preconditioned scaled conjugate gradient (PSCG) [80] is extended for unsupervised learning, where the gradient is approximated by MCMC sampling. The F-scores are between 67.3% and 79.2%.

Haghighi and Klein [81] broke away from the trend in recent work of complex discourse modeling. They instead built the system upon modularized syntactic and semantic constraint filters, akin to early heuristic systems but using sophisticated compatibility filters learned from large unlabeled corpora or motivated by linguistic theories. For example, contrary to commonly used heuristics that determine appositives by matching a "NP, NP" pattern, their filter requires information from a parse tree. The best F-scores ranged from 81.9% using MUC-6, 80.8% using $\boldsymbol{B}^3$, to 73.3% using CEAF.

Pronoun resolution in an unsupervised setting has received its own attention as is in the supervised realm. Cherry and Bergsma [82] first generated a list of candidate antecedents from the parse tree and the (third person) pronoun's context, and then fed this list to an EM algorithm to induce a distribution over the list that maximizes the observed data. Charniak and Elsner [83] improved upon their method by jointly determining anaphoricity and antecedent for pronouns of all three persons.

6. Evaluation metrics

Because the coreference relation is reflexive, symmetric, and transitive[9], simply comparing pairwise predictions cannot appropriately reveal the underlying structure. Consider, for instance, a document with three markables m_1, m_2, and m_3, which are grouped into two pairs $\langle m_1,m_2\rangle$ and $\langle m_2,m_3\rangle$. It would be fastidious to penalize a system that predicted $\langle m_1,m_2\rangle$ and $\langle m_1,m_3\rangle$. A more gentle and intuitive scoring metric would take into account the fact that the gold standard and the system output partition the set of markables in the same way by computing the transitive closures. The first widely adopted evaluation scheme [84] that addresses this issue, developed for the coreference task in MUC-6, was based on the idea of comparing equivalence classes, rather than the links themselves. They follow the standard IR metrics of precision, recall and F-score. Recall errors are calculated by the least number of links that need to be added to the system output in order to align with the gold standard. Precision errors are obtained by reversing the roles of system output and gold standard.

Bagga and Baldwin [85] identified two shortcomings of the MUC-6 score. One is that it does not give credit for recognizing singletons, chains with only one markable. The other is that it intrinsically favors larger chains, as it does not differentiate errors that result from incorrect larger chains from those that only place a smaller number of markables into an incorrect chain. Their revision, the B^3 metric, computes recall and precision by looking at the presence or absence of entities relative to each of the other entities in the equivalence classes. The algorithm proceeds by first computing precision and recall for each markable, and takes the weighted sum.

However, B^3 score has its own drawback. As pointed out by Luo [86], it can give counter-intuitive scores for certain system output. Luo [86] attributed the problem to the process of intersecting the gold standard and system output, during which an entity can be used more than once. Therefore, they proposed a new metric, Constrained Entity-Alignment F-Measure (CEAF). At the heart of the metric is the one-to-one alignment of the gold standard chains and system output chains, which solves the problem of reusing entities in B^3.

Metrics are also borrowed from other domains, although they are rarely reported in the literature for coreference systems. For example, Popescu-Belis et al. [87] adopted κ [88] that is commonly used for measuring inter-coder agreement in annotation tasks. Krippendorff's α [89] is another coefficient developed in the content analysis domain to measure the agreement between observers, coders, judges, raters, or measuring instruments. It is also used in measuring inter-coder agreement [90] for anaphoric annotations, and could be used as a coreference evaluation metric.

Although many systems report multiple scores, the state of the art of coreference resolution systems is still difficult to answer. Some systems report performance on MUC corpus, while others report that on ACE corpus, and not all metrics are reported. Assumptions about gold standard phrasal boundaries further complicate the problem. To properly isolate the coreference task from the underlying named entity recognition task, gold standard NEs are necessary, while in practice, these NEs usually have to be automatically generated.

[9] Strictly speaking, coreference relation is not reflexive (the interpretation of a markable does not depend on itself) or symmetric (the antecedent's interpretation does not depend on the anaphor's), and is only "weakly" transitive (if m_2 depends on m_1 to be interpreted and m_3 depends on m_2, m_3 does not necessarily depend on m_1 for its interpretation). However, for most applications, grouping markables that refer to the same entity is more important than identifying the detailed one-to-one relationship between an antecedent and an anaphor.

Considering these complications, we note that the scores of best performing systems in the general English domain range from 0.7 to 0.8.

7. Discussion

Coreference resolution is a crucial component in an information retrieval system. To link related information in a coherent manner, an information retrieval or extraction system requires coreference resolution. Since MUC and ACE initiated shared tasks, and made their corpora available, supervised machine learning techniques are the predominant efforts in the general NLP community. Two classes of approaches gradually emerged—binary classification (followed by clustering) and ranking. Recently, unsupervised models have also been proposed. Although most unsupervised systems still fall short to the best supervised systems, the performance reported in Poon and Domingos [78] is only a few percentage points lower, and on some data sets even outperforms them. As the community starts to explore more complex systems recently, efforts are also made to refine detailed aspects of a comprehensive system, such as more sophisticated methods to utilize semantic information [91,92].

Research on coreference resolution has largely been focused on pronominals and NPs, the general community has lately started to undertake an even more ambitious coreference task, namely resolving coreference not limited to NPs, as demonstrated in the 2011 Conference on Natural Language Learning (CoNLL) shared task.[10] The main addition in this task is that verbs are also candidates for coreference. In Example 9, resolving the coreference relation between the verb phrase “fell down” and the noun phrase “the incident” would help link the minor hemorrhage to its cause.

However, NLP in the clinical domain has not seen developments in this important area. This can be partly attributed to the lack of sharable annotated clinical corpora. In the general domain, the two corpora have driven much of the progress. Almost all systems discussed in Section 4 are built with either the MUC or the ACE corpus. All systems in Section 5 (except for those pronoun resolution systems) are evaluated on one or both of the two corpora. In contrast, clinical narratives annotated for coreference were not available to the general research community until very recently with the release of ODIE corpus [24]. This corpus is part of the 2011 i2b2/VA shared task, whose first track[11] focuses exclusively on coreference in the clinical domain. By Fall of 2011, the available annotated clinical data combining ODIE and i2b2 will approximate 500,000 words, a size suitable for machine learning.

Similar problems plague the biomedical domain, a related but different source of text. Segura-Bedmar et al. [93] noted that due to the lack of annotated resources, early approaches were mostly based on heuristics [33,94]. Nevertheless, a few more recent coreference resolution systems [57,95-98] were built on machine learning techniques, leveraging the GENIA corpus [58] and the vast amount of data, albeit unannotated, in MEDLINE.

As noted in one of the systems [98] that the biomedical texts differ from newswire, we can hypothesize that the clinical text manifests its own patterns as well, as clinical text are generally cursory, not edited, and abound with idiosyncratic shorthands. This further exemplifies the importance of a corpus in the clinical domain. The i2b2 NLP shared task[12]

[10] http://conll.bbn.com/.

[11] https://www.i2b2.org/NLP/Coreference/.

[12] https://www.i2b2.org/NLP/.

in the clinical domain has released a set of de-identified clinical notes, albeit with annotations geared towards medical knowledge mining, instead of NLP tasks.

Notwithstanding the idiosyncrasies of the clinical narrative discussed here and in Section 1.3, the methods reviewed in this paper can be applied to clinical texts. Although some features such as animacy are not likely to contribute much to the performance, most of the features listed in Table 2 can be easily adapted to the clinical texts. For instance, the WordNet sense can be substituted with UMLS semantic types.

However, the coreference resolution task in clinical texts is intertwined with other attributes of NEs. For example, sentences like Example 10 are common in a clinical report. The two mentions of “pain” may appear to be coreferential. However, it is questionable to assert that the second mention, a worsening pain, is the same as the original mention. Instead, the text may be describing multiple episodes of pain that have occurred over some period of time.

Negation is another example of an named entity’s attributes that may complicate coreference resolution. The two mentions of “PNA” in Example 11 pose a more challenging question—does the initially negated PNA refer to the PNA that was later discovered?

8. Conclusions

Our goal was to review recent advances in general purpose coreference resolution to lay the foundation for methodologies in the clinical domain, facilitated by the availability of a shared lexical resource of gold standard coreference annotations, the Ontology Development and Information Extraction (ODIE) corpus. We reviewed coreference resolution approaches in the general English domain and contrasted them with those in the clinical domain and the related biomedical domain. The methods already developed in the general domain need to be explored for portability to the clinical domain. One key step towards that is the availability of shared annotated resources, which are becoming available and undoubtedly will advance methodologies for information extraction from the clinical narrative.

Acknowledgments

The work was funded by Grant R01 CA127979 (The ODIE tool-kit—software for Ontology Development and Information Extraction).

References

[1]. Fiszman, M.; Haug Peter, J.; Frederick, PR. Automatic extraction of PIOPED interpretations from ventilation/perfusion lung scan reports; Proc AMIA Symp; 1998; p. 860-4.

[2]. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. Medex: a medication information extraction system for clinical narratives. J Am Med Inform Assoc. 2010; 17(1):19–24. doi:10.1197/jamia.M3378. [PubMed: 20064797]

[3]. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity, and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak. 2010; 6:30. doi:10.1186/1472-6947-6-30. [PubMed: 16872495]

[4]. Li, L.; Chase, HS.; Patel, CO.; Friedman, C.; Weng, C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study; Proceedings of the AMIA annual symposium; 2008; p. 404-8.

[5]. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-Treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res. 2010; 62(8): 1120–7. doi:10.1002/acr.20184.

[6]. Kullo IJ, Fan J, Pathak J, Savova GK, Ali Z, Chute CG. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. J Am Med Inform Assoc. 2010; 17(5):568–74. doi:10.1136/jamia.2010.004366. [PubMed: 20819866]

[7]. Savova, GK.; Fan, J.; Ye, Z.; Murphy, SP.; Zheng, J.; Chute, CG., et al. Discovering peripheral arterial disease cases from radiology notes using natural language processing; AMIA Annu Symp Proc; 2010; p. 722-6.

[8]. Garla V, Lo Re III V, Dorey-Stein Z, Kidwai F, Scotch M, Womack J, et al. The Yale cTAKES extensions for document classification: architecture and application. J Am Med Inform Assoc. in press. doi:10.1136/amiajnl-2011-000093.

[9]. Hirst, G. Lecture notes in computer science. Vol. 119. Springer-Verlag; Berlin Heidelberg: 1981. Anaphora in natural language understanding: a survey.

[10]. Coreference task definition; Proceedings of the 6th message understanding conference; 1995; p. 333-44.

[11]. Hirschman, L.; Chinchor, N. Coreference task definition; Proceedings of the 7th message understanding conference; 1997;

[12]. Mitkov, R. Based on the COLING'98/ACL'98 tutorial on anaphora resolution. Anaphora resolution: the state of the art; 1999.

[13]. Ng, V. Supervised noun phrase coreference research: the first fifteen years; Proceedings of the 48th annual meeting of the association for computational linguistics; 2010; p. 1396-411.

[14]. Coden AR, Pakhomov SV, Ando RK, Duffy PH, Chute CG. Domain-specific language models and lexicons for tagging. J Biomed Inform. 2005; 38(6):422–30. doi:10.1016/j.jbi.2005.02.009. [PubMed: 16337567]

[15]. Meystre SM, Savova GK, Kipper-Schuler Karin C, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. IMIA Yearbook 2008: Access Health Inform. 2008; 1:128–44.

[16]. Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A general natural language text processor for clinical radiology. J Am Med Inform Assoc. 1994; 1(2):161–74. doi:10.1136/jamia.1994.95236146. [PubMed: 7719797]

[17]. Haug, P.; Koehler, S.; Lau, LM.; Wang, P.; Rocha, R.; Huff, S. A natural language understanding system combining syntactic and semantic techniques; Proc Annu Symp Comput Appl Med Care; 1994; p. 247-51.

[18]. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc. 2000; 7(6):593–604. doi:10.1136/jamia.2000.0070593. [PubMed: 11062233]

[19]. Hahn U, Romacker M, Schulz S. medSynDiKATe—a natural language system for the extraction of medical information from findings reports. Int J Med Inform. 2002; 67(1-3):63–74. doi:10.1016/S1386-5056(02)00053-9. [PubMed: 12460632]

[20]. Friedman, C. A broad-coverage natural language processing system; Proceedings of AMIA symposium; 2000; p. 270-4.

[21]. Goryachev, S.; Sordo, M.; Zeng, QT. A suite of natural language processing tools developed for the i2b2 project; AMIA Annu Symp Proc; 2006; p. 931

[22]. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010; 17(5):507–13. doi:10.1136/jamia.2009.001560. [PubMed: 20819853]

[23]. Hahn, U.; Romacker, M.; Schulz, S. medSynDiKATe—design considerations for an ontology-based medical text understanding system; Proc AMIA Symp; 2000; p. 330-4.

[24]. Savova GK, Chapman WW, Zheng J, Crowley RS. Anaphoric relations in the clinical narrative: corpus creation. J Am Med Inform Assoc. 2011; 18(4):459–65. doi:10.1136/amiajnl-2011-000108. [PubMed: 21459927]

[25]. Coden A, Savova GK, Sominsky I, Tanenblatt M, Masanz JJ, Schuler K, et al. Automatically extracting cancer disease characteristics from pathology reports into a cancer disease knowledge

model. J Biomed Inform. 2009; 42(5):937–49. doi:10.1016/j.jbi.2008.12.005. [PubMed: 19135551]

[26]. Fritz, A.; Percy, C.; Jack, A.; Shanmugarathan, K.; Sobin, L.; Parkin, DM., et al., editors. International classification of diseases for oncology. World Health Organization; 2000.

[27]. Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Roberts I, et al. Building a semantically annotated corpus of clinical text. J Biomed Inform. 2009; 42(5):950–66. doi: 10.1016/j.jbi.2008.12.013. [PubMed: 19535011]

[28]. Sidner CL. Focusing for interpretation of pronouns. Am J Comput Linguist. 1981; 7(4):217–31.

[29]. Rich, E.; LuperFoy, S. An architecture for anaphora resolution; Proceedings of the second conference on applied natural language processing; Austin (Texas, USA): Association for Computational Linguistics. 1988; p. 18-24.doi:10.3115/ 974235.974239

[30]. Hobbs JR. Resolving pronoun references. Lingua. 1978; 44(4):311–38. doi: 10.1016/0024-3841(78)90006-2.

[31]. Lappin S, Leass HJ. An algorithm for pronominal anaphora resolution. Comput Linguist. 1994; 20(4):535–62.

[32]. Kennedy, C.; Boguraev, B. Anaphora for everyone: pronominal anaphora resolution without a parser; Proceedings of the 16th international conference on computational linguistics; 1996;

[33]. Castaño, J.; Zhang, J.; Pustejovsky, J. Anaphora resolution in biomedical literature; Proceedings of the international symposium on reference resolution for NLP; Alicante, Spain. 2002;

[34]. McCarthy, JF.; Lehnert, WG. Using decision trees for coreference resolution; Proceedings of the fourteenth international joint conference on artificial intelligence (IJCAI'95); Montreal, Quebec. 1995; p. 1050-5.

[35]. Soon WM, Ng HT, Lim DCY. A machine learning approach to coreference resolution of noun phrases. Comput Linguist. 2001; 27(4):521–44.

[36]. Ng, V.; Cardie, C. Improving machine learning approaches to coreference resolution; Proceedings of the 40th annual meeting of the association for computational linguistics; Philadelphia (PA). 2002; p. 104-11.doi:10.3115/1073083.1073102

[37]. Ng, V.; Cardie, C. Combining sample selection and error-driven pruning for machine learning of coreference rules; Proceedings of the 2002 conference on empirical methods in natural language processing; 2002; p. 55-62.

[38]. Yang, X.; Zhou, G.; Su, J.; Tan, CL. Coreference resolution using competition learning approach; Proceedings of the 41st annual meeting of the association for computational linguistics; Association for Computational Linguistics. 2003; p. 176-83.doi:10.3115/1075096.1075119

[39]. Harabagiu, SM.; Bunescu, R.; Maiorano, SJ. Text and knowledge mining for coreference resolution; Second meeting of the North American chapter of the association for computational linguistics; 2001; doi:10.3115/1073336.1073344

[40]. Uryupina, O. Linguistically motivated sample selection for coreference resolution; Proceedings of DAARC; Furnas, Portugal. 2004;

[41]. Luo, X.; Ittycheriah, A.; Jing, H.; Kambhatla, N.; Roukos, S. A mention-synchronous coreference resolution algorithm based on the bell tree; Proceedings of the 42nd meeting of the association for computational linguistics (ACL'04); Barcelona, Spain. 2004; p. 135-42.doi: 10.3115/1218955.1218973

[42]. Ng, V. Machine learning for coreference resolution: from local classification to global ranking; Proceedings of the 43rd annual meeting of the association for computational linguistics; 2005; p. 157-64.

[43]. Denis, P.; Baldridge, J. Specialized models and ranking for coreference resolution; Proceedings of the 2008 conference on empirical methods in natural language processing; Honolulu (HI): Association for Computational Linguistics. 2008; p. 660-9.

[44]. Wagner RA, Fischer MJ. The string-to-string correction problem. J ACM. 1974; 21(1):168–73. doi:10.1145/321796.321811.

[45]. Denis, P.; Baldridge, J. Joint determination of anaphoricity and coreference resolution using integer programming; Human language technologies 2007: the conference of the North American chapter of the association for computational linguistics; proceedings of the main conference; Rochester (New York): Association for Computational Linguistics. 2007; p. 236-43.

[46]. Yang, X.; Su, J.; Tan, CL. Kernel-based pronoun resolution with structured syntactic knowledge; Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics; Sydney, Australia. 2006; p. 41-8.doi: 10.3115/1220175.1220181

[47]. Miller GA. WordNet: a lexical database for English. Commun ACM. 1995; 38(11):39–41. doi: 10.1145/219717.219748.

[48]. Ng, V. Shallow semantics for coreference resolution; Proceedings of the 20th international joint conference on artifical intelligence; Hyderabad, India. 2007; p. 1689-94.

[49]. Bengtson, E.; Roth, D. Understanding the value of features for coreference resolution; EMNLP 2008: proceedings of the conference on empirical methods in natural language processing; Honolulu, HI. 2008; p. 294-303.

[50]. Kehler, A. Probabilistic coreference in information extraction; Proceedings of the second conference on empirical methods in natural language processing (EMNLP-97); 1997; p. 163-73.

[51]. Morton, TS. Coreference for NLP applications; Proceedings of the 38th annual meeting of the association for computational linguistics; 2000; p. 173-80.doi:10.3115/1075218.1075241

[52]. Culotta, A.; Wick, M.; McCallum, A. First-order probabilistic models for coreference resolution; Human language technology conference of the North American chapter of the association of computational linguistics (HLT/NAACL); Rochester (NY): Association for Computational Linguistics. 2007; p. 81-8.

[53]. Uryupina, O. Corry: a system for coreference resolution; Proceedings of the 5th international workshop on semantic evaluation; 2010; p. 100-3.

[54]. Ge, N.; Hale, J.; Charniak, E. A statistical approach to anaphora resolution; Proceedings of the sixth workshop on very large corpora; 1998; p. 161-71.

[55]. McCallum, A.; Wellner, B. In: Saul, LK.; Weiss, Y.; Bottou, L., editors. Conditional models of identity uncertainty with application to noun coreference; Advances in neural information processing systems; Cambridge (MA): MIT Press. 2004; p. 905-91.

[56]. Ng, V.; Cardie, C. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution; Proceedings of the 19th international conference on computational linguistics; Taipei. 2002c; doi:10.3115/1072228.1072367

[57]. Yang, X.; Su, J.; Zhou, G.; Tan, CL. An NP-cluster based approach to coreference resolution; COLING ’04: Proceedings of the 20th international conference on computational linguistics; Geneva (Switzerland). 2004; p. 226-32.doi:10.3115/1220355.1220388

[58]. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics. 2003; 19(Suppl 1):i180–2. doi:10.1093/bioinformatics/btg1023. [PubMed: 12855455]

[59]. Yang, X.; Su, J.; Lang, J.; Tan, CL.; Liu, T.; Li, S. An entity-mention model for coreference resolution with inductive logic programming; Proceedings of ACL-08: HLT; Columbus (OH): Association for Computational Linguistics. 2008; p. 843-51.

[60]. Nicolae, C.; Nicolae, G. BestCut: a graph algorithm for coreference resolution; Proceedings of the 2006 conference on empirical methods in natural language processing; Sydney (Australia): Association for Computational Linguistics. 2006; p. 275-83.

[61]. Rahman, A.; Ng, V. Supervised models for coreference resolution; Proceedings of the 2009 conference on empirical methods in natural language processing; Singapore. 2009; p. 968-77.

[62]. Finley, T.; Joachims, T. Supervised clustering with support vector machines; International conference on machine learning (ICML); 2005; p. 217-24.

[63]. Bansal N, Blum A, Chawla S. Correlation clustering. Mach Learn. 2004; 56(1–3):89–113. doi: 10.1023/B:MACH.0000033116.57574.95.

[64]. Daumé, H., III; Marcu, D. A large-scale exploration of effective global features for a joint entity detection and tracking model; Proceedings of human language technology conference and conference on empirical methods in natural language processing; Vancouver (British Columbia, Canada): Association for Computational Linguistics. 2005; p. 97-104.

[65]. Daumé, H., III; Marcu, D. Learning as search optimization: approximate large margin methods for structured prediction; International conference on machine learning (ICML); Bonn, Germany. 2005; p. 169-76.doi:10.1145/ 1102351.1102373

[66]. Cohen, WW. Fast effective rule induction; Proceedings of the 12th international conference on machine learning; 1995; p. 115-23.

[67]. Evans R. Applying machine learning toward an automatic classification of It. J Lit Linguist Comput. 2001; 16(1):45–57. doi:10.1093/llc/16.1.45.

[68]. Bean, DL.; Riloff, E. Corpus-based identification of non-anaphoric noun phrases; Proceedings of the 37th annual meeting of the association for computational linguistics; College Park (Maryland, USA): Association for Computational Linguistics. 1999; p. 373-80.doi: 10.3115/1034678.1034737

[69]. Ng, V. Learning noun phrase anaphoricity to improve conference resolution: issues in representation and optimization; Proceedings of the 42nd meeting of the association for computational linguistics (ACL'04); Barcelona, Spain. 2004; p. 151-8.doi: 10.3115/1218955.1218975

[70]. Ariel M. Referring and accessibility. J Linguist. 1988; 24(1):65–87. doi:10.1017/S0022226700011567.

[71]. Gundel JK, Hedberg N, Zacharski R. Cognitive status and the form of referring expressions in discourse. Language. 1993; 69(2):274–307.

[72]. Strube, M.; Rapp, S.; Müller, C. The influence of minimum edit distance on reference resolution; Proceedings of the 2002 conference on empirical methods in natural language processing. Association for Computational Linguistics; 2002; p. 312-9.doi:10.3115/1118693.1118733

[73]. Zelenko, D.; Aone, C.; Tibbetts, J. Coreference resolution for information extraction; ACL 2004: workshop on reference resolution and its applications; 2004; p. 24-31.

[74]. Bergsma, S.; Lin, D.; Goebel, R. Distributional identification of non-referential pronouns; Proceedings of ACL-08: HLT; Columbus (Ohio): Association for Computational Linguistics. 2008; p. 10-8.

[75]. Cardie, C.; Wagstaff, K. Noun phrase coreference as clustering; Proceedings of the join SIGDAT conference on empirical methods in natural language processing and very large Corpora; 1999; p. 82-99.

[76]. Haghighi, A.; Klein, D. Unsupervised coreference resolution in a nonparametric bayesian model; Proceedings of the 45th annual meeting of the association of computational linguistics; Prague, Czech Republic: Association for Computational Linguistics. 2007; p. 848-55.

[77]. Ng, V. Unsupervised models for coreference resolution; Proceedings of the 2008 conference on empirical methods in natural language processing; Honolulu, Hawaii. 2008; p. 640-9.

[78]. Poon, H.; Domingos, P. Joint unsupervised coreference resolution with Markov Logic; Proceedings of the 2008 conference on empirical methods in natural language processing; Honolulu (HI): Association for Computational Linguistics. 2008; p. 650-9.

[79]. Richardson M, Domingos P. Markov logic networks. Mach learn. 2006; 62(1–2):107–36. doi: 10.1007/s10994-006-5833-1.

[80]. Lowd, D.; Domingos, P. Efficient weight learning for markov logic networks; Proceedings of the 11th European conference on principles and practices of knowledge discovery in databases (PKDD); 2007; p. 200-11.

[81]. Haghighi, A.; Klein, D. Simple coreference resolution with rich syntactic and semantic features; Proceedings of the 2009 conference on empirical methods in natural language processing. Singapore: Association for Computational Linguistics; 2009; p. 1152-61.

[82]. Cherry, C.; Bergsma, S. An expectation maximization approach to pronoun resolution; Proceedings of the ninth conference on computational natural language learning (CoNLL-2005); Ann Arbor (Michigan): Association for Computational Linguistics. 2005; p. 88-95.

[83]. Charniak, E.; Elsner, M. EM works for pronoun anaphora resolution; Proceedings of the 12th Conference of the European chapter of the ACL (EACL 2009); Athens, Greece: Association for Computational Linguistics. 2009; p. 148-56.

[84]. Vilain, M.; Burger, J.; Aberdeen, J.; Connolly, D.; Hirschman, L. A model-theoretic coreference scoring scheme; MUC6'95: Proceedings of the 6th conference on message understanding; Morristown (NJ, USA): Association for Computational Linguistics. 1995; p. 45-52.doi: 10.3115/1072399.1072405

[85]. Bagga, A.; Baldwin, B. Algorithms for scoring coreference chains; The first international conference on language resources and evaluation workshop on linguistics coreference; 1998;

[86]. Luo, X. On coreference resolution performance metrics; Proceedings of the conference on human language technology and empirical methods in natural language processing; Vancouver (BC): Association for Computational Linguistics. 2005; p. 25-32.doi:10.3115/1220575.1220579

[87]. Popescu-Belis, A.; Rigouste, L.; Salmon-Alt, S.; Romary, L. Online evaluation of coreference resolution; Proceedings of 4th international conference on language resources and evaluation (LREC 2004); Lisbon, Portugal. 2004; p. 1507-10.

[88]. Carletta J. Assessing agreement on classification tasks: The kappa statistic. Comput Linguist. 1996; 22(2):249–54.

[89]. Krippendorff K. Estimating the reliability, systematic error and random error of interval data. Educ Psychol Measur. 1970; 30(1):61–70. doi:10.1177/001316447003000105.

[90]. Poesio, M.; Artstein, R. The reliability of anaphoric annotation, reconsidered: taking ambiguity into account; Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky; Ann Arbor (MI): Association for Computational Linguistics. 2005; p. 76-83.

[91]. Yang, X.; Su, J. Coreference resolution using semantic relatedness information from automatically discovered patterns; Proceedings of the 45th annual meeting of the association of computational linguistics; 2007; p. 528-35.

[92]. Huang, Z.; Zeng, G.; Xu, W.; Celikyilmaz, A. Accurate semantic class classifier for coreference resolution; Proceedings of the 2009 conference on empirical methods in natural language processing; 2009; p. 1232-40.

[93]. Segura-Bedmar I, Crespo M, Pablo-Sánchez C, Martínez P. Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. BMC Bioinform. 2010; 11(Suppl 2):S1. doi:10.1186/1471-2105-11-S2-S1.

[94]. Kim, JJ.; Park, JC.; Bio, AR. anaphora resolution for relating protein names to proteome database entries. In: Harabagiu, S.; Farwell, D., editors. ACL 2004: Workshop on Reference Resolution and its Applications; Barcelona (Spain): Association for Computational Linguistics. 2004; p. 79-86.

[95]. Liang, T.; Lin, YH. Anaphora resolution for biomedical literature by exploiting multiple resources. In: Dale, R.; Wong, KF.; Su, J.; Kwong, OY., editors. Natural language processing—IJCNLP 2005. Lecture Notes in Artificial Intelligence; Springer-Verlag. 2005; p. 742-53.

[96]. Gasperin, C. Semi-supervised anaphora resolution in biomedical texts; Proceedings of the HLT-NAACL BioNLP workshop on linking natural language and biology; New York (New York): Association for Computational Linguistics. 2006; p. 96-103.

[97]. Su, J.; Yang, X.; Hong, H.; Tateisi, Y.; Tsujii, J. In: Ashburner, M.; Leser, U.; Rebholz-Schuhmann, D., editors. Coreference resolution in biomedical texts: a machine learning approach; Ontologies and text mining for life sciences: current status and future perspectives. No. 08131 in Dagstuhl seminar proceedings; Dagstuhl (Germany): Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, Germany. 2008;

[98]. Gasperin, C.; Briscoe, T. Statistical anaphora resolution in biomedical texts; Proceedings of the 22nd international conference on computational linguistics (Coling 2008); Manchester (UK): Coling 2008 Organizing Committee. 2008; p. 257-64.

Table 1

Examples.

1	… he continues to have *significant pain in the shoulder*. … He uses Tylenol … to deal with *his discomfort*.
2	Small focus of *invasive grade 2* (*of 4*) *adenocarcinoma*$_{m1}$ arising in association with *a serrated adenoma*$_{m2}$ with …. The focus of *adenocarcinoma*$_{m3}$ shows invasion into superficial submucosa and is located approximately …. Lateral margins are involved by *adenoma*$_{m4}$ but are negative for *carcinoma*$_{m5}$.
3	… had periods of *chest discomfort* while at rest … complains of some *mild chest tightness* that may last 30-minutes.
4	we briefly discussed *back extensor strengthening exercises* for osteoporosis, and I think she would be an excellent candidate for a home program of *that variety*.
5	The patient presents with *gastrointestinal symptoms* including nausea, vomiting. The patient has had *symptoms* for 10 days. In fact, is having *that problem* since early pregnancy but worst since 10 days.
6	Her pain control appears to be adequate with the *Tramadol* increased to q.i.d. dosing.
7	She had an *arthrogram* in 2030. We have *those films*. *They* show the capsule is tight, and *they* show the cartilage of the glenoid is present.
8	I would support continuing *speech therapy*$_{m1}$ for his *speech deficit m*$_2$. … He had a … *stroke*$_{m3}$ and resultant *aphasia*$_{\mathrm{m4}}$ after *the eventm*$_5$.
9	Patient *fell down* a flight of stairs. *The incident* caused minor hemorrhage.
10	… presents with progressive right sided chest *pain*… The *pain* is worsened with deep inspiration or movement.
11	… the CXR was without any evidence of *PNA*,… subsequently received a CTA to evaluate for PE which revealed multifocal bilateral *PNA*…

Table 2

Common machine learning features in binary classification. m_1 denotes antecedent, m_2 denotes anaphor, and m denotes either. Note that there is much variation in the implementation of these features, and a system listed next to a feature does not necessarily use its exact form. For instance, whereas many systems used gender and number agreements as two different features, Ng and Cardie [36] used a composite of them. Moreover, definitions of certain features are not clear, and in fact can be different in different systems. Finally, this list is not comprehensive, as some systems use domain-specific features, and yet others do not report details [45]. For example, McCarthy and Lehnert [34] used a feature that encodes whether one markable refers to a joint venture child.

Features	Example or explanation	Systems
m_1 and m_2 string match	$m_1 = m_2 =$ "cancer"	[36,41,40,42,43]
$m_1(m_2)$ is substring of $m_2(m_1)$	"tumor" and "the tumor"	[36,41–43]
actual strings	"tumor"	[41]
edit distance between m_1 and m_2	Wagner and Fischer [44]	[41]
$m_1(m_2)$ spans $m_2(m_1)$	m_1 is embedded in m_2	[36,42]
m is prenominal modifier	"tumor" in "tumor size"	[36,42]
m is a pronominal	"the one"	[35,36,40,42,43]
m is a proper name	"Smith"	[35,36,42,43]
m is a subject	"He" in "He is 30 yo."	[36,42]
m is definite	"the tumor"	[35,36,42,40,43]
m is indefinite	"tumor"	[43]
m is demonstrative	"this tumor"	[35,42]
m is possessive	"his knee"	[41]
m is reflexive	"himself"	[41]
NP head	"knee" in "his knee"	[43]
number of *m*	single or plural	[35,36,41,40,42,43]
gender of *m*	masculine, feminine	[35,36,41–43]
person of *m*	1st, 2nd, or 3rd	[40,43]
animacy match		[36,42]
m in quoted string	he said, "the pain" worsens	[36,42]
distance between m_1 and m_2	number of words	[34,35,41,42,40,43]
semantic class agreement	both are disorder markables	[35,36,42]
m_2 is appositive of m_1	"Mr. Smith, the patient"	[35,41,40,42]
$m_1(m_2)$ is an alias of $m_2(m_1)$	"paracetamol" and "acetaminophen"	[34,35,42]
$m_1(m_2)$ is an acronym of $m_2(m_1)$	"ms" and "multiple sclerosis"	[41,43]
WordNet sense	meaning from WordNet	[36,40,42,43]
synonym, antonym		[40]
POS tag	adjective	[41,43]

Table 3

First-order logic predicates proposed in Culotta et al. [52] to expand on pairwise features shown in Table 2. *X* and *Y* can be any pairwise feature.

Predicate	**True iff …**
All-*X*	*X* for all possible pairs in the cluster is true
Most-true-*X*	*X* for a majority of pairs in the cluster is true
Most-false-*X*	*X* for a majority of pairs in the cluster is false
All-true	All pairs are predicted to be coreferent
Most-true	Most pairs are predicted to be coreferent
Most-false	Most pairs are predicted to be non-coreferent
Max-true	The maximum pairwise score is above threshold
Min-true	The minimum pairwise score is above threshold
X^Y	Features *X* and *Y* are both true

Table 4

Summary of supervised coreference resolution systems.

Mention-pair systems	Classification	Partitioning	Performance
[34]	Decision tree	-	85.8–86.5
[35]	Decision tree	Closest-first	60.4–62.6
[36]	Decision tree	Best-first	63.4–70.4
[37]	RIPPER [66]	Best-first	63.4–69.5
[40]	RIPPER	Closest-first	48.0–55.3
[48]	Decision tree	Closest-first	62.3–64.2
[53]	SVM	Integer linear programming	57.2–84.6
[49]	Averaged perceptron	Best-first	78.3–81.8
Entity-mention systems	Classification	Partitioning	Performance
[57]	Decision tree	Closest- and best-first	81.2–81.7
[59]	Inductive logic programming	Best-first	60.1–63.5
[41]	Maximum entropy	Bell tree search	72.1–85.7
[60]	Maximum entropy	Graph cutting	41.2–89.6
Ranking systems	Algorithm		Performance
[38]	Decision tree		60.2–71.3
[43]	Maximum entropy		67–71.6
[52]	Maximum entropy		69.2–79.3
[61]	SVM		59.5–76.0
Other systems	Algorithm		Performance
[62]	Supervised clustering with SVM		
[55]	Graph partitioning		60.83–73.42
[64]	LaSO [65]		76.7–89.2
[42]	SVM ranker		54.7–69.3