Information Fusion 59 (2020) 139–162

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Anaphora and coreference resolution: A review

Rhea Sukthanker[a], Soujanya Poria[b], Erik Cambria[c,*], Ramkumar Thirunavukarasu[d]

[a] Department of Computer Science, ETH Zurich, Switzerland
[b] Information Systems Technology and Design, SUTD, Singapore
[c] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[d] School of Information Technology and Engineering, VIT University, Vellore, India

ARTICLE INFO

Keywords:
Coreference resolution
Anaphora resolution
Natural language processing
Sentiment analysis
Deep learning

ABSTRACT

Coreference resolution aims at resolving repeated references to an object in a document and forms a core component of natural language processing (NLP) research. When used as a component in the processing pipeline of other NLP fields like machine translation, sentiment analysis, paraphrase detection, and summarization, coreference resolution has a potential to highly improve accuracy. A direction of research closely related to coreference resolution is anaphora resolution. Existing literature is often ambiguous in its usage of these terms and often uses them interchangeably. Through this review article, we clarify the scope of these two tasks. We also carry out a detailed analysis of the datasets, evaluation metrics and research methods that have been adopted to tackle these NLP problems. This survey is motivated by the aim of providing readers with a clear understanding of what constitutes these two tasks in NLP research and their related issues.

1. Introduction

A discourse is a collocated group of sentences which convey a clear understanding only when read together. The etymology of anaphora is *ana* (Greek for back) and *pheri* (Greek for to bear), which in simple terms means repetition. In computational linguistics, anaphora is typically defined as references to items mentioned earlier in a discourse or “pointing back” reference as described by Mitkov [93]. The most prevalent type of anaphora in natural language is the pronominal anaphora [73]. Coreference, as the term suggests refers to words or phrases referring to a single unique entity (or union of entities) in an operating environment. Anaphoric and co-referent mentions themselves form a subset of the broader term “discourse parsing” [135], which is crucial for full text understanding. In spite of having a rich research history in the NLP community, the progress of anaphora resolution (AR) research has not been as rapid as some other subfields of NLP because of the challenges involved in this task. Some applications of this task in NLP span crucial fields like sentiment analysis [14], summarization [136], machine translation [118], question answering [20], etc. AR can be seen as a tool to confer these fields with the ability to expand their scope from intra-sentential level to inter-sentential level.

This paper aims at providing the reader with a coherent and holistic overview of AR and coreference resolution (CR) problems in NLP. These fields have seen a consistent and steady development, starting with the earlier rule-based systems [65,73] to the recent deep learning based methodologies [27,28,79,157,163]. Though there have been some thorough and intuitive surveys, the most significant ones are by Mitkov [93] for AR and [101,102] for CR. The detailed survey on AR by Mitkov [93] provides an exhaustive overview of the syntactic constraints and important AR algorithms. It also analyzes the applications of AR in other related NLP fields. The most recent survey by Ng [102] targets the research advances in the related field of CR delineating the mention-pair, entity-mention and mention-ranking models proposed till date. Both of these surveys are a great resource to gain a deeper understanding of research methodologies which have been attempted earlier for AR and CR.

Another closely related area is the area of event CR, which is arguably more challenging than entity CR. The difficulty stems from the fact that it relies on several components of information extraction pipeline which yield particularly noisy results. Focusing on entity CR is a way of restricting the task of CR by allowing an NP to only co-refer with an NP but in some cases it can co-refer with an event too. Take for example the sentence “The citizens streamed the streets of the city shouting slogans (1). The riots (2) lasted for a week.”. Here, (2) co-refers with (1), but (1) is not necessarily an NP. Most of the research methods discussed in this paper are centered around entity CR. There have also been attempts to jointly address event and entity CR [77,81]. Since a detailed survey of event CR methodologies is beyond the scope of this paper, we refer readers to the works of [7,77,81,143] for this topic.

* Corresponding author.
E-mail addresses: srhea@student.ethz.ch (R. Sukthanker), sporia@sutd.edu.sg (S. Poria), cambria@ntu.edu.sg (E. Cambria), ramkumar.thirunavukarasu@vit.ac.in (R. Thirunavukarasu).

https://doi.org/10.1016/j.inffus.2020.01.010
Received 19 May 2019; Received in revised form 19 January 2020; Accepted 30 January 2020
Available online 1 February 2020

The advent of deep neural networks in NLP has demonstrated performance strides in most of its sub-fields, including POS tagging [31], social data analysis [108], dependency parsing [23], personality detection [90], etc. and this is no different to the field of CR. In fact, this paper is fueled by the necessity for a detailed analysis of new approaches pertaining to this field. Our focus throughout the review is mainly on the English language. Here, we build on earlier surveys by delineating the pioneering research methodologies proposed for these two very closely related, yet significantly different, fields of research. Often the proposed methodologies differ in the evaluation metrics adopted, thus making the comparison of their performance a major challenge. To this end, we provide a comprehensive section on evaluation metrics, with the aim of establishing well-defined standards for comparison. Another motivation for this survey is the requirement to establish standard datasets and open-source toolkits for researchers and commercial users.

AR and CR have seen a shifting trend, from methods completely dependent on hand-crafted features to deep learning based approaches, which attempt to learn feature representations and are loosely based on hand-engineered features. As this trend looks very promising, we have discussed it in the comparison section. Another important topic is the type of references that can occur in sentences and the constraints to be applied to identify possible co-referring entities. Though state-of-the-art approaches have demonstrated a significant margin of improvement from the earlier ones, some rare types of references have gone unnoticed.

This field has faced a long history of debate with regards to comparison of different types of approaches, appropriate metrics for evaluation, right preprocessing tools, etc. One hot topic of debate pertaining to CR, which we cover, is whether induction of commonsense knowledge aids the resolution process. Through this survey, we also analyze the application of CR to other NLP tasks, e.g., sentiment analysis [22]. Finally, this survey aims to form the building blocks for the reader to better understand this exciting field of research.

In the following sections, we dive into the depths of the tasks of AR and CR. We start with describing and comparing the two tasks, thus laying the foundation for the survey. We then delineate the types of references in the English language and the constraints which typically need to be accounted for resolution of references. We then look into evaluation metrics, thus establishing well-defined comparison standards. Going forward, we look into the datasets annotated for these tasks along with the methodologies and off-the-shelf tools available to tackle this area of research. Finally, we conclude the paper with a discussion on the issues and controversies faced by these fields.

2. Introduction to anaphora and coreference resolution

AR is an intra-linguistic terminology, which means that it refers to resolving references used within the text with a same sense (i.e., referring to the same entity). Also, these entities are usually present in the text and, hence, the need of world knowledge is minimal. CR, on the other hand, has a much broader scope and is an extra-linguistic terminology. Co-referential terms could have completely different grammatical structure and function (e.g., gender and part of speech) and yet, by definition, they could refer to the same extra linguistic entity. Here, *entity* could be a single object in a world or a group of objects which together form a new single entity. CR treats entities in a way more similar to how we understand discourse, i.e., by treating each entity as a unique entity in real time.

From the above explanation, it may seem that AR is a subset of CR. However, this claim though commonly fails in some cases as stated by Mitkov [94] in his example: *Every speaker had to present his paper*. Here, if "his" and "every speaker" are said to co-refer (i.e., refer to the same entity), the sentence is interpreted as "Every speaker had to present Every speaker's paper" which is obviously not correct. Thus, "his" here is

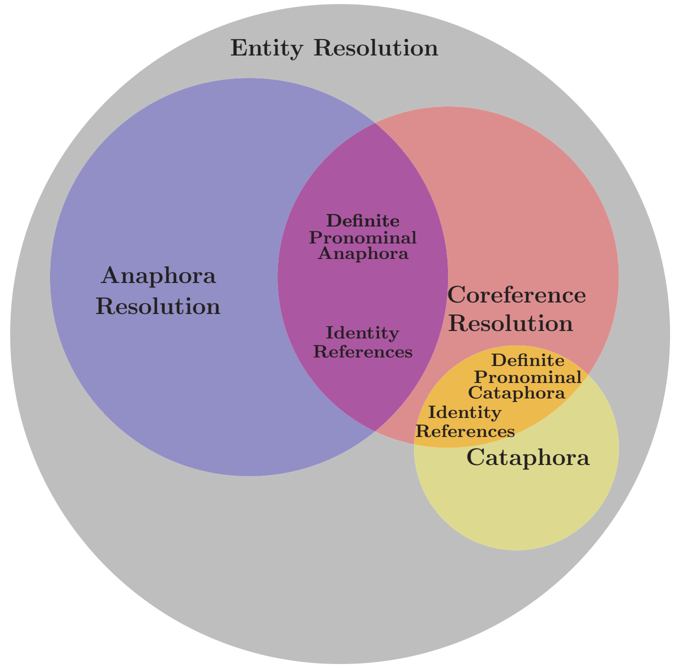

Fig. 1. Venn Diagram of Entity Resolution.

an anaphoric referent and not co-referential, hence demarcating the two very similar but significantly different concepts. This is a typical case of the bound variable problem in entity resolution. Hence, the often made claim that AR is a type of CR, fails in this case.

Some researchers also claim that coreference is a type of AR. However, this can often be seen as a misnomer of the term "anaphora", which clearly refers to something that has occurred earlier in a discourse. CR, on the other hand, spans many fields like AR, cataphora resolution, split antecedent resolution, etc. For example: *If he(1) is unhappy with your work, the CEO(2) will fire you*. Here, the first reference is not anaphoric as it does not have any antecedent, but (1) is clearly co-referent with (2). What we see here is the occurrence of the cataphora phenomenon. Thus, this claim too fails to capture this phenomenon adequately. Though these two concepts have a significant degree of overlap, they are quite different (Fig. 1).

There is a clear need for redefinition of the CR problem. We find that standard datasets like CoNLL 2012 [116] fail to capture the problem to its entirety. To address the fuzziness involved in the terminologies used in entity resolution, we suggest that the datasets created for the task explicitly specify the coreference type they have considered for annotation and the ones they have not. We also insist that future entity resolution (CR, AR, etc.) models also perform exhaustive error analysis and clearly state the types of references which their algorithms fail to resolve. This serves two purposes: firstly, it helps future researchers focus their efforts on specific types of references like co-referent event resolution, which most models fail to resolve, and secondly, this helps in highlighting some issues in the way we currently define the task.

3. Types of references in English

AR is a particularly challenging task because of the different forms the "references" can take. Most AR and CR algorithms face the "coverage" issue. This means that most algorithms are designed to target only specific types of references. Before proceeding to the current state-of-the-art research methodologies proposed for this task, it is necessary to understand the scope of this task to its entirety. In this section, we discuss the different semantic and syntactic forms in which references can occur.

3.1. Zero anaphora

This type of anaphora is particularly common in prose and ornamental English and was first introduced by Fillmore [45]. It is perhaps one of the most involved type of AR task which uses a gap in a phrase or a clause to refer back to its antecedent. For example, in the sentence "*You always have two fears(1): your commitment(2) versus your fear(3)*" phrases (2) and (3) refer back (are anaphoric) to the same phrase (1).

3.2. Demonstratives

This type of reference as explained by Dixon [40] is typically used in contexts when there is a comparison between something that has occurred earlier. For example, in the sentence "*I like this dessert better than that*". Here, *that* is a demonstrative pronoun which refers to something that has been seen before.

3.3. Presuppositions

In this type of references, the pronouns used are commonly singular indefinite pronouns. Here, the resolution is complicated as there is a degree of ambiguity involved in the consideration of the noun phrases (NPs) which the referents can be corresponding to. The projection of presupposition as an AR task was first introduced by Van der Sandt [132]. For example, in the sentence "*If there is anyone(1) who can break the spell, it is you(2)*", the phrase (2) co-refers with (1). Here, the major source of ambiguity is the phrase "anyone".

3.4. Discontinuous sets (split anaphora)

The issue of clause splitting in AR has been delineated by Mitkov [96]. In this type of anaphora, the pronoun may refer to more than one antecedents. Commonly (not always) the pronouns which refer to more than one antecedents are personal pronouns. For example, in the sentence "*Kathrine(1) and Maggie(2) love reading. They(3) are also the members of the reader's club.*", the pronoun (3) refers to Maggie (2) and Katherine (1) together as a single entity (note that here our final entity is a union of two entities). Most of the prominent algorithms in anaphora and CR fail to consider this phenomenon. Later in the paper, we discuss one recent research methodology that specifically focuses on this issue.

3.5. Pronominal anaphora

This is one of the most common and prevalent types of anaphora which occur in day-to-day speech and constitutes a significant portion of the anaphors we commonly see in web data like reviews, blog posts, etc. An example of this type is "*The opposition(1) raised their(2) voices against the newly passed bill*". Here, (2) is the pronominal referent of (1). This type of anaphora initially introduced by Roberts [131] and Heim [64], has been the focus of many papers. The earliest one being the paper of [73] which aimed at pronominal resolution. There exist six types of pronominal anaphors: one anaphora, indefinite pronominal, definite pronominal and adjectival pronominal.

3.5.1. One anaphora

In this type of anaphora, the pronoun"one" is used to refer to the antecedent. This type of anaphora, though not very common, has received sufficient attention from the research community, particularly the machine learning approach by Ng and Cardie [104] which specifically targeted this type. The one anaphora phenomenon can be best illustrated with an example. In the sentence "*Since Samantha has set her eyes on her friend's poodle(1), she wants one (2)*", the phrase (2) refers back to the entity depicted by (1).

3.5.2. Indefinite pronominal

In this reference type, the pronoun refers to an entity or object which is not well-defined or well-specified. An example of this type is "*Many(1) of the jurors(2) held the same opinion*", where (2) refers back to (1), though the exact relation between the referents is ambiguous.

3.5.3. Definite pronominal

This type of reference is definite since it refers to a single unique entity in the universe. For example, in the sentence "*She had seen the car(1) which met with an accident. It(2) was an old white ambassador*", pronoun *it*(2) refers back to entity (1) *the car*.

3.5.4. Adjectival pronominal

In this type of anaphora, there is reference to adjectival form of the entity which has occurred earlier. For example, in the sentence "*A kind stranger(1) returned my wallet. Such people(2) are rare*", (2) refers back to (1). Thus, (1) here is an adjectival form that has been referred to by the anaphor (2). This example also serves to illustrate that adjectival noun forms can also be anaphoric.

3.6. Cataphora

Cataphora as defined by Mitkov et al. [97] is said to be the opposite of anaphora. A cataphoric expression serves to point to entities which may succeed it. The phenomenon of cataphora is most commonly seen in "poetic" English. For example, in the sentence "*If she(1) doesn't show up for the examination even today, chances of Clare(2) clearing this semester are meagre*", (1) refers to an entity that precedes it, i.e., (2). In this paper, do not review techniques for cataphora resolution.

3.7. Inferable or bridging anaphora

Bridging anaphora [66] is perhaps one of the most ambiguous ones. They may not explicitly seem to be pointing to an antecedent but can be said to belong to or refer to an entity mentioned at some point earlier in time. For example, in the sentence "*I was about to buy that exquisite dress(1); just when I noticed a coffee stain on the lace(2)*", the entity that (2) refers to, though not stated explicitly, is entity (1) which can be inferred by their context.

3.8. Non-anaphoric pronominal references

A major issue which is unanimously tackled by all state-of-the-art methods is the identification and elimination of empty referents or referents which potentially do not refer back to any antecedent. The major categories of non-referential usage are: clefts and pleonastic "it".

3.8.1. Clefts

A cleft sentence is a complex sentence (one having a main clause and a dependent clause) that has a meaning that could be expressed by a simple sentence. A cleft sentence typically puts a particular constituent into focus. Clefts were introduced by Atlas and Levinson [2]. For example, in the sentence "it" cleft is: "*It(1) was Tabby who drank the milk*", (1) does not serve to refer to an antecedent but is a potential trap for most AR systems.

3.8.2. Pleonastic "it"

This issue in AR has received a lot of attention and has been delineated by Mitkov [96]. This type of non-anaphoric referent is very common in natural language. For example, in the sentence "*It(1) was raining heavily*", (1) in spite of being a pronoun does not refer to any specific entity. The dummy pleonastic "it" can also be found in the extraposition [55] construct in the English language.

4. Constraints for anaphora and coreference resolution in English

Most proposed approaches in AR and CR are based on some trivial syntactic and semantic constraints. Though all constraints may not be relevant for every type of referent, most methods do apply some if not all of these constraints. Syntactic approaches are solely based on these constraints and exploit them to a large extent for AR. Most statistical and machine learning approaches use these constraints in feature extraction or mention-filtering phase. Recently, there has been a growing trend towards knowledge-poor AR [92]. This mainly aims at reducing the level of dependency on these hand-crafted rules. Also, it is important to understand here that these constraints are not universally acceptable, i.e., some may not hold true across different languages. The constraints below are necessary but not sufficient by themselves to filter out the incorrect references. This section aims at delineating the linguistic constraints for AR.

4.1. Gender agreement

Any co-referring mentions should agree on their gender, i.e., masculine, feminine, neuter or common gender.[1] Gender is a very important constraint in AR as mentioned by Mitkov [96]. Antecedents that do not agree in terms of their gender need not be considered further for evaluation of their correctness. This is one of the crucial constraints which serves to prune the antecedent search space to a large extent. For example, in the sentence "*Tom(1) bought a puppy(2). It(3) is adorable*", on application of this constraint, (1) is eliminated due to gender disagreement with (3), thus culminating in (it=puppy). The question which arises here is what happens when there are multiple antecedents satisfying gender constraint. This brings about the necessity to enforce some other syntactic and semantic constraints.

4.2. Number agreement

A mention may co-refer with another mention if and only if they agree on the basis of their singularity and plurality. This constraint is even incorporated into machine learning systems like [104]. This constraint is necessary but not sufficient and the final resolution may be subject to the application of other constraints. For example, in the sentence "*Fatima and her sisters(1) bought groceries(2) for the week(3). Recently, there has been a huge hike in their(4) prices*", the pronominal reference (4) refers to (2) and (1). Referent (3) is successfully eliminated on the basis of number disagreement.

4.3. Constraints on verbs (selectional constraints)

Human language type-casts certain verbs to certain entity classes. There are certain verbs which occur with only animate or living entities and some others specifically on the inanimate ones. Constraints on verbs have been exploited in many methods like [57]. The sentence "*I sat on the tortoise(1) with a book(2) in my hand, assuming it to be a huge pebble and that's when it(3) wiggled*", for example is very difficult for a computer to interpret. Here, (3) can refer to (2) or (1). The reference (2) should be filtered out here using the animacy constraint. This constraint brings about the necessity to incorporate world knowledge into the system.

4.4. Person agreement

The English language has three persons. The First person refers to the speaker (e.g., I and me), the Second person refers to the one being spoken to (e.g., you and yourself) and the Third person refers to the one being spoken about (e.g., he and them). This feature has been exploited by many approaches like [73]. The co-referent nouns or entities must agree with respect to their person. For example, in the sentence "*Me and you(1) are not siblings. No doubt we(2) are so different from each other*", (2) refers to (1) as they agree with respect to their person. In case the pronoun (2) had been "they" this possible antecedent would have been eliminated.

4.5. Grammatical role

Very often a given sentence can be decomposed to its subject, verb and object part and these roles of the words in the sentence can aid AR as mentioned by Kennedy and Boguraev [68]. Mentions occurring in the subject portion of a sentence are given a higher priority than the mentions in object position. For example, in the sentence "*Kavita(1) loves shopping. She goes shopping with her sister(2) every weekend. She(3) often buys stuff that she may never use*, (3) refers to (1) and not to (2) as (1) being the subject has more priority or salience over (2).

4.6. Recency

As mentioned in [18], recency is an important factor of consideration in AR. Entities introduced recently have more salience than entities which have occurred earlier in a discourse. For example, in the sentence "*I have two dogs. Steve(1), a grey hound, is a guard dog. Bruno(2) who is a Labrador is pampered and lazy. Sally often takes him(3) for a stroll*", (3) may refer to (1) or (2) syntactically. To resolve this ambiguity this constraint gives more salience to (2) over (3) due to the mention's recency.

4.7. Repeated mention

Repeated mention forms a feature of many systems like the statistical method of [48]. Entities which have been introduced repeatedly in the context or have been the main focus or topic of the earlier discourse are given a higher priority than the rest. For example, in the sentence "*Katherine(1) is an orthopaedic surgeon. Yesterday she ran into a patient(2), she had not been in contact with since ages. She(3) was amazed at her speedy recovery*", the referent (3) refers to (1) and not (2) because Katherine (1) here is a mention that has been in focus in prior discourse and, hence, is more salient.

4.8. Discourse structure

The preference of one mention over another can also be due to the structural idiosyncrasies of a discourse like parallelism. These phenomenon are discussed by Carbonell and Brown [18] in their paper and form a crucial component of Centering Theory. For example, in the sentence "*Aryan(1) passed a note to Shibu(2) and Josh(3) gave him(4) a chocolate*", (4) refers to (2) and not (1). Here, *Shibu* is the indirect object of the verb *to pass*,Â and *him* is the indirect object of the verb *toÂ give*. Thus the two verbs are coordinated and hence (3) is resolved to refer to (2). This type of reference can be motivated by observing that (2) is the entity in focus currently and by centering theory [53] it is more likely for the center of attention to stay the same here. Though the occurrence of this type of discourse is tough to spot and disambiguate, if exploited appropriately this can increase the precision factor involved in the CR to a large extent.

4.9. World knowledge

This is the constraint that has a very wide scope and generally cannot be completely incorporated into any system. In spite of this, attempts to incorporate this behavior in CR systems has been made by Rahman and Ng [123]. Though syntax does play a role in CR, to some extent world knowledge does function as a critical indicator. This case is quite different from the selectional constraints mentioned in Section 4.3. Here, the verbs/nouns are often uniquely attributed to a particular object. World knowledge can be broadly categorized as common and commonsense

[1] https://universaldependencies.org/u/feat/Gender.html

knowledge: the former consisting mostly of named entities, the latter consisting of physical and behavioral knowledge about the world and society [15]. One example where commonsense knowledge is needed for disambiguating anaphora is "*We went for dinner and movie. It was delicious!*", where the pronoun refers to the dinner and not the movie. Another example is the sentence "*And once again in the history of tennis FedEx delivers*", where common knowledge needs to be incorporated to disambiguate "FedEx" as the tennis legend "Roger Federer".

5. Evaluation metrics in coreference resolution

There are a number of metrics which have been proposed for the evaluation of CR and AR. Here, we delineate the most standard ones for the CR task. In our opinion, this choice is justified by the fact that current research [79] and datasets available [116] are mainly driven towards CR.

5.1. B^3 (B-Cubed)

This metric proposed by Bagga and Baldwin [3] begins by computing precision and recall for each individual mention and, hence, takes weighted sum of these individual precisions and recalls. Greedy matching is undertaken for evaluation of chain-chain pairings.

$$Recall_i = \frac{\text{number of correct elements in the output chain containing entity i}}{\text{number of elements in the true chain containing entity i}} \tag{1}$$

$$Precision_i = \frac{\text{number of correct elements in the output chain containing entity i}}{\text{number of elements in the output chain containing entity i}} \tag{2}$$

$$Final\,Precision = \sum_{i=1}^{N} w_i * Precision_i \tag{3}$$

$$Final\,Recall = \sum_{i=1}^{N} w_i * Recall_i \tag{4}$$

Where N= Number of entities in the document and w_i is the weight assigned to entity i in the document and $Precision_i$ and $Recall_i$ are the individual precision and recall. Usually the weights are assigned to $1/N$.

5.2. MUC

This metric, proposed during the 6th Message Understanding Conference (MUC) by Vilain et al. [150] considers a cluster of references as linked references, wherein each reference is linked to at most two other references. MUC metric primarily measures the number of link modifications required to make the result-set identical to the truth-set.

$$partition(c, s) = \{s | s \in S \& s \in c \neq \phi\} \tag{5}$$

MUC Precision value is calculated as follows:

$$MUC\,Precision(T, R) = \sum_{r \in R} \frac{|r| - |partition(r, T)|}{|r| - 1} \tag{6}$$

Where, $|partition(r, T)|$ is the number of clusters within truth T that the recall cluster r intersects with. MUC Recall value is calculated as follows:

$$MUC\,Recall(T, R) = \sum_{t \in T} \frac{|t| - |partition(t, R)|}{|t| - 1} \tag{7}$$

Where, $|partition(t, R)|$ represents the number of clusters within the result R that truth set $|t|$ intersects with.

5.3. CEAF

The Constrained Entity Alignment F-measure (CEAF) metric proposed by Luo [82] is used for entity-based similarity identification. It uses similarity measures to first create an optimal mapping between result clusters and truth clusters. Using this mapping, CEAF leverages self-similarity to calculate the precision and recall. Luo [82] proposes four different types of similarity measurements. We use T to represent the key entities and R to represent the response entities. $\phi_1(T, R)$ insists that two entities are the same if all the mentions are the same, $\phi_2(T, R)$ states that two entities are same if they share at least one common mention, ϕ_3(T,R) counts the number of common mentions shared by T and R and ϕ_4(T,R) is the F-measure between T (key entities) and R (response entities). These similarity measures are computed using the following equations:

$$\phi_1(T, R) = \begin{cases} 1 & \text{if R=T} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$\phi_2(T, R) = \begin{cases} 1, & \text{if } R \cap T \neq \phi \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

$$\phi_3(T, R) = |R \cap T| \tag{10}$$

$$\phi_4(T, R) = 2.\frac{|R \cap T|}{|R| + |T|} \tag{11}$$

The function $m(r)$ takes as an input a cluster r and returns the true cluster t that the result cluster r is mapped to with constraint that one cluster can be mapped to at most one result cluster. CEAF Precision and Recall is defined as follows:

$$CEAF_{\phi_i} Precision(T, R) = max_m \frac{\sum_{r \in R} \phi_i(r, m(r))}{\sum_{r \in R} \phi_i(r, r)} \tag{12}$$

$$CEAF_{\phi_i} Recall(T, R) = max_m \frac{\sum_{r \in R} \phi_i(r, m(r))}{\sum_{r \in R} \phi_i(t, t)} \tag{13}$$

5.4. ACE

The ACE evaluation score [41] proposed during the Automatic Content Extraction (ACE) conference is also based on optimal matching between the result and the truth like CEAF. The difference between the two is that ACE's precision and recall is calculated using true positives, false positives, false negatives amongst the predicted co-referent entities. Another difference is that ACE does not normalize its precision and recall values unlike the CEAF metric.

5.5. Conll score

This score is calculated as the average of B^3, MUC and CEAF scores. This is the score used by the CoNLL 2012 shared task by Pradhan et al. [116] which is based on CR in the OntoNotes corpus. Thus, the CoNLL score is calculated as shown in the equation below.

$$CoNLL = \frac{(MUC_{F1} + B^3{}_{F1} + CEAF_{F1})}{3} \tag{14}$$

5.6. BLANC

BLANC [115,127] is a link-based metric that adapts the Rand index [124] CR evaluation. Given that C_k is the set of coreference links in the key entities, C_r is the set of coreference links in the response entities, N_k is the set of non-coreference links in the key entities and N_r is the set of coreference links in the response entities, the BLANC Precision and Recall is calculated as follows. R_c and R_n refer to recall for coreference links and non-coreference links, respectively. Precision is also defined with a similar notation.

$$R_c = \frac{|C_k \cap C_r|}{|C_k|} \tag{15}$$

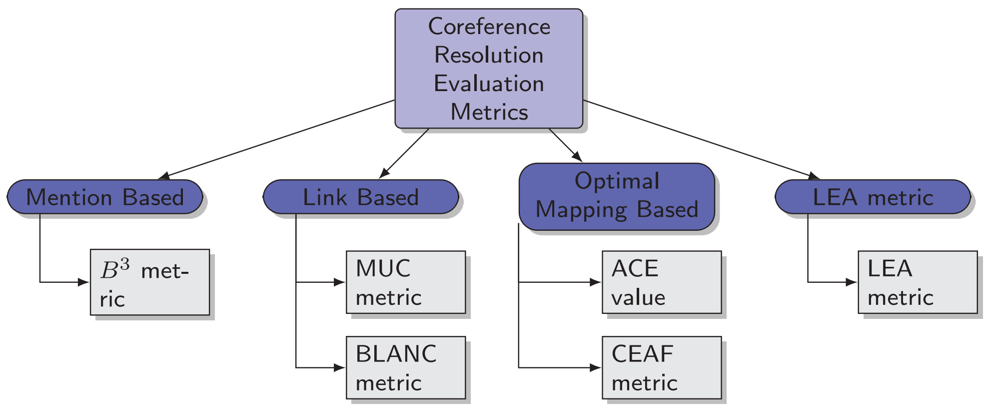

Fig. 2. Evaluation Metrics.

$$P_c = \frac{|C_k \cap C_r|}{|C_r|} \tag{16}$$

$$R_n = \frac{|N_k \cap N_r|}{|N_k|} \tag{17}$$

$$P_n = \frac{|N_k \cap N_r|}{|N_r|} \tag{18}$$

$$Recall = \frac{R_c + R_n}{2} \tag{19}$$

$$Precision = \frac{P_c + P_n}{2} \tag{20}$$

The mention identification effect delineated by Moosavi and Strube [98] highly affects BLANC and, hence, this metric is not widely adopted.

5.7. LEA

Link-based Entity Aware (LEA) metric proposed by Moosavi and Strube [98] aims at overcoming the mention identification effect of the earlier coreference evaluation metrics which makes it impossible to interpret the results properly. LEA considers how important the entity is and how well is it resolved. LEA metric is dependent on two important terminologies which are *importance* and *resolution-score*. Importance is dependent on size of the entity and resolution-score is calculated using link similarity. The link function returns the total number of possible links between n mentions of an entity e.

$$\text{importance}(e_i) = |e_i| \tag{21}$$

$$\text{resolution-score}(k_i) = \sum_{r_j \in R} \frac{link(k_i \cap r_j)}{link(k_i)} \tag{22}$$

$$Recall = \frac{\sum_{k_i \in K} importance(k_i) * \sum_{r_j \in R} \frac{link(k_i \cap r_j)}{link(k_i)}}{\sum_{k_z \in K} importance(k_z)} \tag{23}$$

$$Precision = \frac{\sum_{r_i \in R} importance(r_i) * \sum_{k_j \in K} \frac{link(r_i \cap k_j)}{link(r_i)}}{\sum_{r_z \in R} importance(r_z)} \tag{24}$$

In the above equations, r_i represents the result set and k_i represents the key set or the gold set.

5.8. Comparison of evaluation metrics

The MUC score, which was one of the earliest metric to be proposed for CR, has some drawbacks as pointed out by Luo [82]. Being link-based, MUC score ignores singleton-mention entities, since no link can be found in the entities. It also fails to distinguish the different qualities of system outputs and favors system producing fewer entities. Thus, in some cases MUC score may result in higher F-measure for worse systems. B^3 metric, which was MUC score's successor, aimed at fixing some of the problems associated with the metric. However, an important issue associated with B^3 is that the individual precision and recall involved are calculated by comparing entities containing the mention and, hence, an entity can be used more than once. The BLANC metric [125] is also quite flawed because it considers the non-coreference links which increase with increase in gold mentions, thus giving rise to the mention identification effect. ACE metric, which is very closely related to CEAF, is not very interpretable. CEAF metric solves the interpretability issue of the ACE metric and the drawbacks of MUC F1 score and B^3 F1 score. However, CEAF metric has problems of its own too. As mentioned by Denis and Baldridge [38], CEAF ignores all correct decisions of unaligned response entities that may lead to unreliable results. A recent paper which particularly targets this flaw [98] discusses the issues with the existing metrics and proposes a new LEA metric. Fig. 2 represents the different types of metrics proposed till date.

6. Coreference and anaphora resolution datasets

The datasets predominantly used for the task of CR differ based on a number of factors like their domain, their annotation schemes, the types of references which are labelled etc. Thus, it is crucial to develop a clear understanding of the AR and CR datasets before proceeding to the research methodologies pertaining to the task which use these datasets either for training or for deriving effective rules. Every new dataset for AR and CR was introduced with the aim of addressing the issues with the earlier ones. In this section, we provide a categorization of the datasets by whether they label AR or CR.

6.1. Coreference resolution datasets

Though there are multiple datasets available for CR, there are some major ones which have been widely popular for evaluation purposes. The two main corpora created targeting this task were MUC and OntoNotes. MUC-6 [51] and MUC-7 [25] were typically prepared by human annotators for training, dry run test and formal run test usage. The MUC datasets which were the first corpora of any size for CR, are now hosted by Linguistic Data Consortium. These datasets contain 318 annotated Wall Street Journal (WSJ) Articles mainly based on North American news corpora. The co-referring entities are tagged using SGML tagging based on MUC format. The evaluation on this dataset is carried out using the MUC scoring metric. Now that larger resources containing multi-genre documents are available, these datasets are not widely used any more except for comparison with baselines.

The Task-1 of SemEval 2010 defined by Recasens et al. [127] was CR. This dataset is made freely available for the research community. The annotation format of the SemEval Dataset is similar to the CoNLL dataset and it is derived from the OntoNotes 2.0 corpus. SemEval 2010 CR task can be seen as a predecessor of the CoNLL 2012 shared task. The

Table 1
AR and CR datasets Comparison.

Dataset	Multi-lingual	Multi-Domain	Intra-Document Annotation	Inter-Document Annotation
CoNLL 2012	✓	✓	✓	✗
ECB+	✗	✗	✓	✓
SemEval 2010	✓	✓	✓	✗
ARRAU	✗	✓	✓	✗
CIC	✗	✗	✓	✗
MUC 6 & 7	✗	✗	✓	✗
ParCor	✓	✓	✓	✗
GUM	✗	✓	✓	✗
NP4E	✗	✗	✓	✓
ACE	✓	✓	✓	✗
WikiCoref	✗	✓	✓	✗
GNOME	✗	✓	✓	✗

CoNLL 2012 shared task [116] targeted the modeling of CR for multiple languages. It aimed at classifying mentions into equivalence classes based on the entity they refer to. This task was based on the OntoNotes 5.0 dataset which mainly contains news corpora. This dataset has been widely used recently and is freely available for research purposes.

The GUM corpus [164] is another open-source multilayer corpus of richly annotated web texts. It contains conversational, instructional and news texts. It is annotated in the CoNLL format. The WikiCoref dataset [49], maps the CR task to Wikipedia. The dataset mainly consists of 30 annotated Wikipedia articles. Each annotated entity also provides links to FreeBase knowledge repository for the mentions. The dataset is annotated in OntoNotes schema using the MaxNet tagger and is freely available for download. GUM and WikiCoref are both mainly based on Wikipedia data. These datasets aimed to address two main issues in CR datasets: domain adaptation and world knowledge induction.

6.2. Anaphora resolution datasets

The ACE corpus [41], which was the result of series of evaluations from 2000 to 2008, is labelled for different languages like English, Chinese and Arabic. Though the initial version of this corpus was based on news-wire articles like MUC, later versions also included broadcast conversations, web-log, UseNet and conversational telephonic speech. Thus, this dataset is not genre-specific and is heterogeneous. ACE is mainly annotated for pronominal AR and is evaluated using the ACE score metric.

Unlike the datasets discussed earlier, the GNOME corpus by Poesio [111] contains annotated text from the museum, pharmaceutical and tutorial dialogue domains and, hence, is very useful for cross-domain evaluation of AR and CR algorithms. Since this corpus was mainly developed for study of centering theory, it focuses on “utterance” labelling. GNOME follows a scheme developed by the MUC initiative. It also includes a separate specification of the elements to be marked, as in GNOME paragraphs and sentences can also serve as antecedents of anaphoric expressions. It uses XML as a markup language. The GNOME corpus is not freely distributed. The Anaphora Resolution and Underspecification (ARRAU) corpus [112] is a corpus labelled for anaphoric entities and maintained by LDC (Linguistic Data Consortium). This corpus is a combination of TRAINS [52,63], English Pear [153], RST [19] and GNOME datasets. It is labelled for multi-antecedent anaphora, abstract anaphora, events, actions and plans. It is labelled using the MMAX2 format which uses hierarchical XML files for each document, sentence and markable. This is a multi-genre corpora, based on news corpora, task oriented dialogues, fiction, etc. The ARRAU guidelines were also adapted for the LIVEMEMORIES corpus [128] for AR in the Italian Language.

6.3. Miscellaneous and task-specific datasets

In addition to the above corpora, there were some corpora which were created for task-specific CR. The ParCor dataset by Guillou et al. [54] is mainly for parallel pronoun CR across multiple languages. It is based on the genre of TEDx talks and Bookshop publications. The corpora is annotated using the MMAX2 format. This parallel corpus is available in two languages, German and English and mainly aims at CR for machine translation.

The Character Identification Corpus is a unique corpus by Chen and Choi [24] which contains multi-party conversations (tv show transcripts) labelled with their speakers. This dataset is freely available for research and is annotated using the popular CoNLL format. This dataset like GNOME is extremely useful for cross-domain evaluation. This corpus was introduced mainly for the task of character-linking in multi-party conversations.

The task of event CR has also received significant amount of attention. The NP4E corpora [62] is labelled for corefering events in the texts. This corpus is based on terrorism and security genres and is annotated in the MMAX2 format. Another event coreference dataset is the Event Coreference Bank (ECB +) [33] dataset for topic-based event CR. The dataset is available for download and is annotated according to the ECB + format.

In this article, our main focus is the study of the CR task in English, but it is very interesting to note that there are datasets available to address this issue in multiple languages. The ACE corpora [41] and the CoNLL 2012 [116] shared task in addition to English are also labelled for the Chinese and Arabic languages. The SemEval 2010 Task-1 [127] also provides datasets for CR in Catalan, Dutch, German, Italian and Spanish languages. The ParCor [54] corpus is also labelled for German language. The AnCora-Co [128] corpus has also been labelled for coreference in Spanish and Catalan. We provide the reader with a concise overview of the properties of the datasets previously discussed in Tables 1 and 2.

6.4. Biomedical coreference resolution datasets

The above-mentioned datasets are labelled on multiple text genres. Recently, there has also been a surge of interest in the area of domain-specific CR, particularly biomedical CR. This can be attributed to the BioNLP 2011 [72] CR task which was built on the GENIA corpus and contains PubMed abstracts. There are mainly two lines of research in the biomedical CR task: annotation of abstract and full-text annotations.

In abstract annotation, biomedical abstracts are labelled with coreferent entity types. These datasets mainly use annotation scheme like MUC-7 and restrict to labelling of only biomedical entity types. The MedCo[1] corpus described by Su et al. [139] consists of coreference annotated Medline abstracts from GENIA dataset. The Protein Coreference resolution task was a part of BioNLP 2011 shared task [72]. The dataset for the task was a combination of three resources: MedCo coreference annotation [139], Genia event annotation [71], and Genia Treebank [140] all of which were based on the GENIA corpus by Kim et al. [70]. This task focused on resolution of names of proteins. Medstract is a large corpus of medline abstracts and articles labelled for CR which was introduced by Pustejovsky et al. [119]. The coherence and anaphora module of this dataset focuses on resolution of biologically

Table 2
English Coreference and Anaphora Resolution Datasets.

Dataset	Source Corpora	Statistics	Genre	Annotation Scheme	Availability
CoNLL 2012 [116]	OntoNotes 5.0 Corpus	Train docs:2802 Test docs:348 Dev docs:343 Total docs: 3493	News, conversational telephone speech, web-logs, UseNet newsgroups, talk shows	CoNLL format	Freely available through LDC
ECB+ [33]	Google News	Total docs:982	News	ECB+ format	Freely available
SemEval 2010 [127]	English: OntoNotes 2.0	Train docs:229 Test docs:85 Dev docs:39	News, conversational telephone speech, web-logs, UseNet newsgroups, talk shows	CoNLL formatwq	Freely available through LDC
ARRAU 2 [112]	TRAINS, English Pear, RST, GNOME	Total docs:552	News (RST), task-oriented dialogues (TRAINS), fiction (PEAR) and medical leaflets (GNOME)	MMAX2 format	Available by payment through LDC
CIC [24]	Dialogue Scripts of Friends TV Show (Season 1 and 2), and The Big Bang Theory TV Show (Season 1)	Train docs(Episodes+ Scenes): 478 Dev docs(Episodes+ Scenes):51 Test docs(Episodes+Scenes):77	TV Show Dialogues	CoNLL format	Freely available for download
MUC 6 & 7 [51]	MUC 6:WSJ corpus, MUC 7: NY Times Corpu	MUC 6-Train docs:30, Test docs:30, Total docs:60 MUC 7-Train docs:30, Test docs:20, Total docs:50	News	MUC SGML tagging format	Available by payment through LDC
ParCor [54]	Multilingual SMT Corpora	Total docs:19	EU Bookshop and TED Talks	MMAX2	Freely Available
GUM [164]	Wikinews, WikiHow, WikiVoyage, Reddit, Wikipedia	Total docs:101	News (narrative) Interview (conversational) How-to (instructional) Travel guide (informative), Academic Writing, Biographies, Fiction, Forum Discussions	Richly annotated with multiple layers of annotation like RST, CoNLL, WebAnno, ISO date/time, Dependencies etc	Freely Available
NP4E [62]	Reuters Corpus	Total docs: (NP+Event Coreference) 104+20=124	News in domain of Terrorism/Security	Available in NP4E defined annotation and MMAX format	Freely Available
ACE 2007 [41]	News articles from: New York Times, Cable News Network, etc.	Total docs: 599	Weblogs, Broadcast Conversation, Broadcast News, News Groups	ACE format	Available by payment through LDC
WikiCoref [49]	Wikipedia	Total docs:30	People, Organization, Human made Object, Occupation	CoNLL format	Freely Available
GNOME corpus [111]	Museum:ILEX, SOLE corpora Pharmaceuticals: ICONOCLAST corpora, and Dialogues: Sherlock corpus	Total docs:5	Museum, Pharmaceuticals, Tutorial Dialogues	GNOME format	Not Available

Table 3
Comparison of Biomedical coreference datasets.

Dataset	Type	Statistics	Annotation	Availability
MEDSTRACT	Abstract Annotation	100 abstracts	MUCCS	publicly available
MEDCo-A	Abstract Annotation	1999 abstracts	MUCCS	publicly available
MEDCo-B	Full-Text Annotation	43 full texts	MUCCS	currently unavailable
FlySlip	Full-Text Annotation	5 full texts	FlySlip scheme	publicly available
CRAFT	Full-Text Annotation	97 full texts	OntoNotes	currently unavailable
DrugNERAr	Full-Text Annotation	49 full texts	MUCCS	publicly available
HANNAPIN	Full-Text Annotation	20 full texts	MEDCo-scheme	publicly available

relevant sortal terms (proteins, genes as well as pronominal anaphors). It is mainly concerned with two types of anaphora namely pronominal and sortal anaphora. DrugNerAR corpus proposed by Segura-Bedmar et al. [133] aims at resolving anaphora for extraction drug-drug interactions in pharmacological literature. It is derived from the DrugBank corpus which contains 4900 drug entries. This corpus was created by extracting 49 structured, plain unstructured and plain documents which were randomly taken from field interactions and, hence, annotated for nominal and pronominal anaphora.

There are a lot of benefits associated with using full-text instead of abstracts for biomedical CR. Though such fully annotated biomedical texts are not very accessible, there are three very interesting projects which aim at creating this type of corpora. The Corlando Richly Annotated Full-Text or CRAFT corpus [29] contains 97 full-length open access biomedical journal articles that have been annotated both semantically and syntactically to serve as a resource for the BioNLP community. Unlike the other corpora created for CR in biomedical literature, this corpus is drawn from diverse biomedical disciplines and are marked up to their entirety. The FlySlip corpus [47] was introduced with the aim of addressing the issues associated with the earlier BioNLP Corpora which mainly considered only short abstracts. Since anaphora is a phenomenon that develops through a text, this paper posited that short abstracts are not he best resources to work with. The domain of this corpora is fruit fly genomics and it labels direct and indirect sortal anaphora types. The HANNAPIN corpus [5] was a successor of CRAFT corpus which also annotates full biomedical articles for CR. The annotated 20 full-text covers several semantic types like proteins, enzymes, cell lines and pathogens, diseases, organisms, etc. This resource is openly available for researchers. We provide a summary of the biomedical datasets in Table 3.

7. Reference resolution algorithms

7.1. Rule-based resolution

Reference resolution task in NLP has been widely considered as a task which inevitably depends on some hand-crafted rules. These rules are based on syntactic and semantic features of the text under consideration. Which features aid resolution and which do not has been a constant topic of debate. There have also been studies conducted specifically targeting this issue [8,99]. While most of the earlier AR and CR algorithms were dependent on a set of rich hand-crafted rules (often knowledge intensive) and, hence, were knowledge-rich, there were others that aimed at minimizing this dependency.

Hobb's naïve algorithm [65] was one of the first algorithm to tackle AR. This algorithm used a rule-based, left to right breadth-first traversal of the syntactic parse tree of a sentence to search for an antecedent. Hobb's algorithm also used world knowledge based selectional constraints for antecedent elimination. The rules and selectional constraints were used to prune the antecedent search space till the algorithm converged to a single antecedent. This algorithm was manually evaluated on different corpora like fiction and non-fiction books and news magazines.

Another knowledge-rich algorithm was the Lappin and Leass algorithm [73] for pronominal AR. This algorithm was based on the salience assignment principle. This algorithm maintained a discourse model consisting of all potential antecedent references corresponding to a particular anaphor. Each antecedent was assigned a salience value based on a number of features. The salience categories were recency, subject emphasis, existential emphasis, accusative emphasis, indirect object emphasis, non-adverbial emphasis and head noun emphasis. The strategy followed here was to penalize or reward an antecedent based on its syntactic features. The algorithm started with generation of a list of possible antecedents extracted using the syntactic and semantic constraints mentioned earlier. Then, a salience value was assigned to each antecedent. The salience was calculated as a sum over all the predetermined salience values corresponding to the salience category satisfied. The antecedent with the maximum salience value was proposed as the appropriate antecedent. The Lappin and Leass algorithm also incorporated a signal attenuation mechanism wherein the influence or salience of an antecedent was halved on propagation to next sentence in a discourse and was evaluated on a dataset consisting of five computer science manuals.

The earliest attempt at exploiting discourse properties for pronoun resolution was the BFP algorithm [12]. This algorithm motivated the centering theory. The centering theory [53] was a novel algorithm used to explain phenomenon like anaphora and coreference using discourse structure. In centering theory, the center was defined as an entity referred to in the text which linked multiple "utterances" or sentences in a discourse. Forward looking centers were defined as set of centers that were referred to in an utterance. The backward looking center was defined as a single center belonging to the intersection of the sets of forward centers of the current and the preceding utterance. This algorithm started with creation of all possible anchors, i.e., pairs of forward centers and backward entities. The ordering of the centers was done according to their prominence and their position in the utterance. The backward looking center was defined as the current topic and the preferred center was the potential new topic. The three major phases in center identification were: center continuation, where same center was continued for the next sentence in discourse, center retaining, wherein there was a possible indication for shift of the center and center-shifting wherein there was a complete shift in the center involved. As summarized by Kibble [69] there were two key rules governing centering theory. The **Rule 1** stated that the center of attention was the entity that was most likely to be pronominalized and **Rule 2** stated that there was a preference given to keep the same entity as the center of attention. Apart from these rules various discourse filters were also applied to filter out good and bad anchors and the remaining good ones were ranked according to their transition type. The centering algorithm was evaluated on Hobb's datasets and some other Human-Keyword task oriented databases. There were many modifications proposed on centering theory and the most significant one was the Left Right Centering theory [141,142]. This was based on the observation in Psycholinguistic research that listeners attempted to resolve an anaphor as soon as they heard it. LRC [141] first attempted to find the antecedent in the current utterance itself and if this does not work it proceeds to process the previous utterances in a left to right fashion. Another modification on LRC, i.e., LRC-F [142] also encoded information about the current subject into the centering theory.

Though most of the rule-based algorithms were knowledge-rich, there were some [4,57,59,75,165] that aimed at reducing the level of dependency of rules on external knowledge. These were categorized as the "knowledge-poor algorithms". CogNIAC [4] was a high precision coreference resolver with limited resources. This early method moved a step closer to how human beings resolve references. Take, for example, the sentence: *Charles (1) went to the concert with Ron (2) and he hurt his (3) knee on the way back.* Here, the resolution of (3) is an intricate task for a human being due to inevitable requirement of knowledge beyond the discourse. Thus, CogNIAC was based on the simple but effective assumption that there existed a sub class of anaphora that did not require general purpose reasoning. Thus, if an anaphoric reference required external world resources in its resolution CogNIAC simply did not attempt its resolution. Here, CogNIAC could be considered to be analogous to a human who recognizes knowledge intensive resolutions and makes a decision on when to attempt resolution. CogNIAC was evaluated on multiple datasets like narratives and newspaper articles and in scenarios with almost no linguistic preprocessing to partial parsing. The core rules defining CogNIAC were picking a unique or single existent antecedent in current or prior discourse, the nearest antecedent for a reflexive anaphor, picking exact prior or current string match for possessive pronoun, etc. Adhering to the these core rules or presuppositions, the CogNIAC's algorithm proceeded to resolve pronouns from left to right in the given text. Rules were followed in an orderly fashion and once a given rule was satisfied and antecedent match occurred no further rules are attempted. On the other hand, if none of the rules were satisfied CogNIAC left the anaphor unresolved. Two additional constraints were deployed during the evaluation phase of CogNIAC. These two constraints were picking the backward center which is also the antecedent as the target solution and the second one was picking the most recent potential antecedent in the text. CogNIAC was evaluated on multiple datasets like narratives and newspaper articles.

Apart from the methods discussed earlier which were a combination of salience, syntactic, semantic and discourse constraints, attempts have also been made to induce world knowledge into the CR systems. The COCKTAIL system [60], basically a blend of multiple rules, was one such system which took a knowledge-based approach to mine coreference rules. It used WordNet for semantic consistency evidence and was based on structural coherence and cohesion principles. It was evaluated on the standard MUC 6 CR dataset.

Another rule-based algorithm which took a knowledge-based approach for pronominal AR was the rule-based algorithm by Liang and Wu [80] for automatic pronominal AR. In this algorithm, WordNet ontology and heuristic rules were deployed to develop an engine for both intra-sentential and inter-sentential antecedent resolution. This algorithm started with parsing each sentence in the text, POS tagging and lemmatizing it. These linguistic features were stored in an internal data structure. This global data structure was appended with some other features like base nouns, number agreement, person name identification, gender, animacy, etc. This model also constructed a finite state machine with the aim of identifying the NPs. The parsed sentence was then sequentially checked for anaphoric references and pleonastic it occurrences. The remaining mentions were considered as possible candidates for antecedents and were heuristically evaluated using a scoring function. The toolkit was extensively evaluated on reportage, editorials, reviews, religion, fiction, etc.

As the research in CR started shifting towards machine learning algorithms which used classification and ranking, it slowly became clear that to beat the machine learning systems, rules had to be ordered according to their importance. A rule-based CR baseline which gained wide acclaim was the deterministic CR system by Haghini and Klein (H and K model) [57], who proposed a strong baseline by modularizing syntactic, semantic and discourse constraints. In spite of its simplicity it outperformed all the unsupervised and most of the supervised algorithms proposed till then. This algorithm first used a module to extract syntactic paths from mentions to antecedents using a syntactic parser. It then proceeded by eliminating some paths based on deterministic constraints. After this, another module was used to evaluate the semantic compatibility of headwords and individual names. Compatibility decisions were made from compatibility lists extracted from corpora. The final step was the elimination of incompatible antecedents and selection of the remaining antecedents so as to minimize the tree distance. This algorithm was evaluated on multiple versions of the ACE corpus and the MUC-6 dataset and achieved significant improvements in accuracy.

The H and K model [57] motivated the use of "successive approximations" or multiple hierarchical sieves for CR. The current version of Stanford CoreNLP deterministic CR system is a product of extensive investigations conducted on deciding the precise rules to govern the task of CR. This system was an outcome of three widely acclaimed papers [75,76,121]. Though rule-based systems have lost their popularity in favor of deep learning algorithms, it is very interesting to understand how this multi-sieve based approach for CR improved over time. The work of [121] was motivated by the hypothesis that a single function over a set of constraints or features did not suffice for CR as lower precision features could often overwhelm higher precision features. This multi-sieve approach proposed a CR architecture based on a sieve that applied tiers of deterministic rules ordered from high precision to lowest precision one be one. Each sieve built on the result of the previous cluster output. The sieve architecture guaranteed that the important constraints were given higher precedence. This algorithm had two phases. The first one was the mention processing phase wherein the mentions were extracted, sorted and pruned by application of multiple constraints. The second phase was the multi-pass sieve phase which used multiple passes like string match, head match, precise constraints like appositives, shared features like gender, animacy, number, etc. This system was evaluated on the same datasets as the H and K model [57] and outperformed most of the baselines.

An extension of the multi-sieve approach [121] was presented at the CoNLL 2011 shared task [117]. The major modifications made to the earlier system were addition of five more sieves, a mention detection module at the beginning and, finally, a post-processing module at the end to provide the result in OntoNotes format. This system was ranked first in both the open and closed tracks of the task. A more detailed report and more extensive evaluation of this system was also reported by Lee et al. [75], who delineated the precise sieves applied using an easy to understand and intuitive example. Like the earlier system [121], this approach also incorporated shared global entity-level information like gender, number and animacy into the system to aid CR. Fig. 3 shows the composition of different sieves used in this deterministic system.

The shifting trend of CR research from rule-based systems to deep learning systems has come at the cost of loss of the ability of the CR systems to adapt to different coreference phenomenon and border definitions, when there is no access to large training data in the desired target scheme [165]. A recent rule-based algorithm [165] also used dependency syntax as input. It aimed at targeting the coreference types which were not annotated by the CoNLL 2012 shared task like cataphora, compound modifier, i-within-i, etc. This system was called Xrenner and was evaluated on two very different corpora, i.e., the GUM corpus and the WSJ. It used semantic and syntactic constraints and rules for antecedent filtering. Xrenner was compared with other well-known rule-based systems like Stanford CoreNLP [75] and Berkeley systems [42] on the datasets mentioned earlier and outperformed all of them. Xrenner raised a very important question of the domain-adaptation problem of learning-based systems.

7.2. Comparative analysis of rule-based systems

Hobb's algorithm [65] was a syntax-based algorithm while centering theory [53] was discourse-based algorithm. Lappin and Leass [73] algorithm, on the other hand, can be seen as a hybrid since it was both syntax- and discourse-based. In addition, it also made use of knowledge resources and morphological and semantic information to rank the pos-

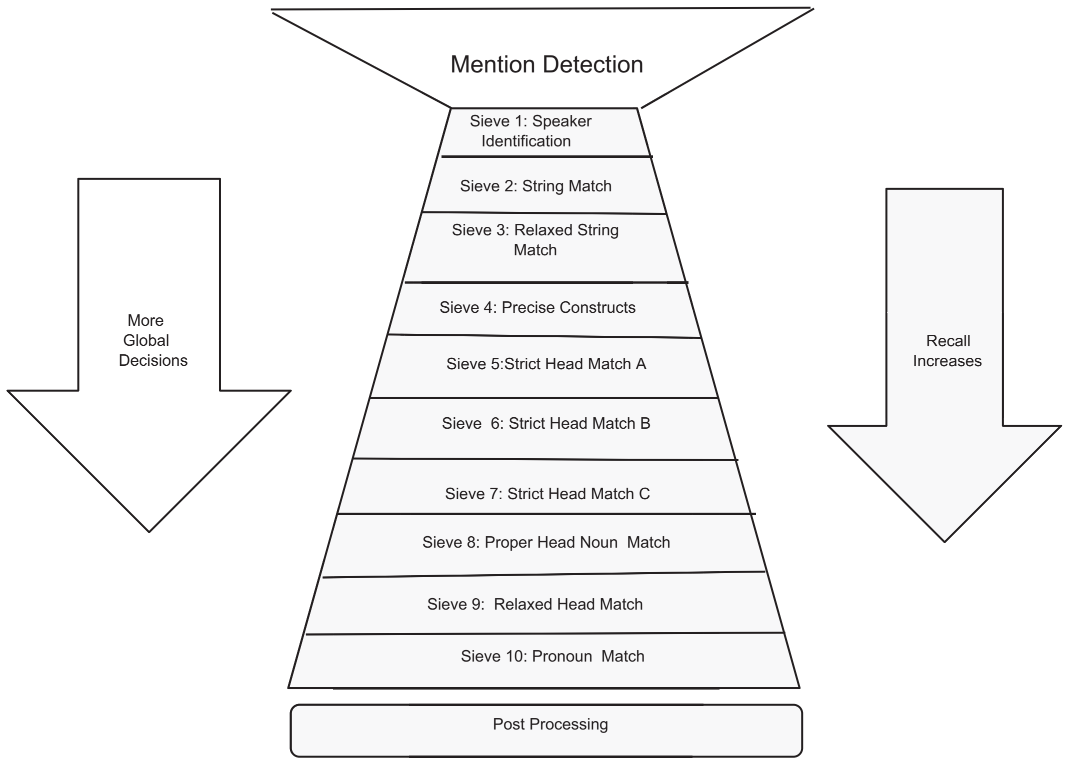

Fig. 3. Coreference Resolution Sieve [75].

sible antecedents. These three algorithms were amongst the earliest algorithms for AR and, hence, the evaluation metrics and datasets used for their evaluation were not the standardized ones. This makes comparison of these algorithms with the recent rule-based algorithms extremely difficult. Also, Hobb's algorithm [65] was hand-evaluated and, hence, was prone to human errors. The datasets used for evaluation of this algorithm also raise a concern. As pointed out by contemporaries, in most cases the resolution of the antecedent was trivial. Another issue with the algorithm is that it was based on the assumption that the correct syntax parse of the sentence always exists (which is often not true). Nonetheless, this algorithm is still highly regarded as a strong baseline given its simplicity and ease of implementation.

The Lappin and Leass algorithm [73] which is still highly regarded in AR research also had some drawbacks. First is that Lappin and Leass algorithm was mainly aimed only at pronominal AR which only forms a small subset of the AR task. Another drawback is that the Lappin and Leass algorithm is highly knowledge driven. This dependency on knowledge resources can become very problematic especially when the required knowledge resources are not accessible. Another loophole of this algorithm is the weight assignment scheme for the different salience categories. These weights are decided by extensive corpus experiments. Hence, the fundamental question which arises here is are these values corpus-dependent. The weight initialization stage can be very problematic when trying to adapt this algorithms to other corpora. The Lappin and Leass algorithm (RAP) was compared with Hobb's algorithm for intra-sentential and intra-sentential case on a dataset of computer manuals. Hobb's algorithm outperformed the RAP algorithm for the inter-sentential case (87% vs 74%) while the RAP algorithm outperformed Hobb's algorithm for the intra sentential case (89% vs 81%). Overall, RAP algorithm outperformed Hobb's algorithm by 4%. In spite of Lappin and Leass algorithm's success, it is important to bear in mind that this algorithm was tested on the same genre as its development set, while the genre used by Hobb for the development of his own algorithm was very different from the test set. High Precision systems like CogNIAC [4] aimed at circumscribing the heavy dependency of the RAP algorithm on external knowledge resources by taking a knowledge-minimalistic approach. CogNIAC achieved performance at-par with Hobb's algorithm in the "resolve-all" setting. In spite of this CogNIAC had issues of its own. Its rules were defined only for specific reference types and it was mainly useful for systems which required high precision resolution at the cost of a low recall. As a result of this, the defined rules performed well on the narratives dataset but CogNIAC failed to meet a high precision when evaluated on the MUC-6 corpus.

The centering theory [53,152] was a discourse-based algorithm that took a psycholinguistic approach to AR. Centering theory was an attractive algorithm for researchers mainly because the discourse information it requires could be obtained from structural properties of utterances alone. Thus, eliminating the need for any extra-linguistic semantic information. One possible disadvantage of CT was its preference for inter-sentential references as compared to intra-sentential references. In some ways we can even consider that Lappin and Leass algorithm incorporated centering theory's discourse constraint and modified it by assigning weights to these discourse phenomenon. A manual comparison of Hobb's algorithm [65] and CT-based algorithm by Walker [151] showed that the two performed equally over a fictional dataset of 100 utterances, but Hobb's algorithm outperformed CT for news paper articles domain (89% vs 79%) and task domain (51% vs 49%). In spite of this spurge in interest in this field with the methods discussed earlier, it is important to note one thing. The evaluation standards of these algorithms were very inconsistent [95] and this slowly started to change with the evaluation guidelines laid by MUC [51], ACE [41] and CoNLL [117] corpora.

Another widely accepted and extensively evaluated rule-based system was the coreference system by Haghini and Klein [57]. This system was evaluated on multiple standard datasets like MUC and ACE corpora. This simple but effective algorithm was purely based on syntax and had well-defined antecedent pruning rules. Instead of weighting the salience categories like Lappin and Leass, this algorithm defined rules which were successively applied starting with the most important ones. This algorithm formed the first effective baseline comparison of rule-based CR approaches. Its main strength was its simplicity and

Table 4
Rule-based Algorithms.

Algorithm	Dataset	Evaluation Metric	Metric Value	Algorithm Rule Types
[65]	Fictional, Non-fictional, Books, Magazines, Part of Brown Corpus	Hobb's metric*	88.3% (without selectional constraints) 91.7% (with selectional constraints)	Syntax-based rules + Selectional rules
[73]	Five Computer Science Manuals	Hobb's metric	74% (inter-sentential) 89% (intra-sentential)	Hybrid of: Syntax Rules + Discourse Rules + Morphological + Semantic
[152]	2 of the fiction and non fiction books same as Hobb's + 5 Human-keyword and task-oriented and task oriented databased	Hobb's metric	Overall: 77.6%	Discourse-based rules and constraints
[4]	Narratives MUC-6	Precision and Recall	P:92% R:64% P:73% R:75%	Discourse rules+Syntax rules
[80]	Random Texts from Brown Corpora	Hoobs's metric	Overall:77%	Semantic constraints + Discourse constraints + Syntactic Constraints
[57]	ACE 2004 Roth-dev ACE 2004 Culotta-test MUC-6 Test ACE 2004 nwire	MUC, B^3, CEAF (F1 values)	MUC:75.9, B^3:77.9,CEAF:72.5 MUC:79.6, B^3:79,CEAF:73.3 MUC:81.9, B^3:75.0,CEAF:72 MUC:76.5, B^3:76.9, CEAF:71.5	Syntactic rules + Semantic rules
[121]	ACE 2004 Roth-dev	MUC, B^3	MUC: 78.6, B^3:80.5	Syntactic rules + Semantic rules(minimal)
	ACE 2004 Culotta-test MUC-6 Test ACE 2004 nwire		MUC:75.8, B^3:80.4 MUC:77.7, B^3:73.2 MUC:78.1, B^3:78.9	
[75]	ACE 2004 Culotta-test ACE 2004 nwire MUC6-Test	MUC, B^3	MUC:75.9, B^3:81 MUC:79.6, B^3:80.2 MUC:78.4, B^3:74.4	Syntactic rules+Semantic rules
	CoNLL 2012	MUC, B^3, CEAF, CoNLL	MUC:63.72, B^3:52.08, CEAF:47.79, CoNLL:60.13	
[165]	GUM corpus	MUC, B^3, CEAF, CoNLL	MUC:55.95, B^3:49.09, CEAFe: 44.47, CoNLL:49.84	Syntactic Rules
	Wall Street Journal Corpus		MUC:49.23, B^3:41.52, CEAFe:41.13, CoNLL: 43.96	

* Hobb's metric $= \frac{\text{Number of correct resolutions}}{\text{Total No. of Resolutions attempted}}$

effectiveness. The idea of defining rules was further developed and delineated more intuitively using a novel sieve architecture for CR [121]. Over time, there were a couple of additions and modifications made to this architecture to improve its performance [75], result of which is the current version of the best performing rule-based system of Stanford CoreNLP. This coreference system is extremely modular and new coreference models can be easily incorporated into it. Overall, we observe that the rules and constraints deployed became even more fine-grained as the CR research took pace. This was mainly because the focus of the task started to shift towards CR which has a much broader scope than AR. We provide a summary of the rule-based approaches in Table 4.

7.3. Statistical and machine learning based resolution

The field of reference resolution underwent a shift during the late nineties from heuristic- and rule-based approaches to learning-based approaches. Some of the early learning-based and probabilistic approaches for AR used decision trees [1], genetic algorithms [95,97] and Bayesian rule [48]. These approaches set the foundation for the learning-based approaches for reference resolution which improved successively over time and, finally, outperformed the rule-based algorithms. This shift was mainly because of the availability of tagged coreference corpora like MUC and ACE corpora. The research community of CR expanded from linguists to machine learning enthusiasts. Learning-based coreference models can be classified into three broad categories of mention-pair, entity-mention and ranking model.

The mention-pair model treated coreference as a collection of pairwise links. It used a classifier to make a decision whether two NPs are co-referent. This stage was followed by the stage of reconciling the links using methods like greedy partitioning or clustering to create an NP partition. This idea was first proposed for pronoun resolution [1,89] in the early nineties using the decision tree classifier [120] and is still regarded as a simple but very effective model. The mention pair model had three main phases each of which acquired significant research attention. It is important to note here that the training of the mention classification and clustering phase is independent and improvement in the performance of one stage did not necessarily imply improvement in accuracy of the other.

The first phase of the mention-pair model was the creation of training instances. Since most entities in the text were non-coreferent, the aim of training instance creation was to reduce the skewness involved in the training samples. The most popular algorithm for mention instance creation was the Soon et al. 's heuristic mention creation method [134]. Soon's method created a positive instance between a NP *A1* and its closest preceding antecedent *A2* and a negative instance by pairing *A1* with each of the NPs intervening between *A1* and *A2*. It only considered annotated NPs for instance creation. A modification on this approach [104] enforced another constraint that a positive instance between a non-pronominal instance *A1* and antecedent *A2* could only be created if *A2* was non-pronominal too. Other modifications on Soon's instance creation [138,160] used number, gender agreement, distance features for pruning of incorrect instances. There have also been some mention creation systems [59,103] which learnt a set of rules with the aim of excluding the hard training instances whose resolution was difficult even for a human being.

The second phase of mention-pair models was the training of a classifier. Decision trees and random forests were widely used as classifiers [1,78,89] for CR. In addition, statistical learners [9,48], memory learners like Timbl [34] and rule-based learners [30] were also widely popular.

The next phase of the mention pair model was the phase of generating an NP partition. Once the model was trained on an annotated corpus it could be tested on a test-set to obtain the coreference chains.

Multiple clustering techniques were deployed to tackle this task. Some of the most prominent ones were best-first clustering [104], closest-first clustering [134], correlational clustering [88], Bell Tree beam search [82] and graph partitioning algorithms [87,105]. In the closest first clustering [134] all possible mentions before the mention under consideration were processed from right to left, processing the nearest antecedent first. Whenever the binary classifier returned true the two references were linked together and the further antecedents were not processed. Further, the references could be clustered using transitivity between the mentions. A modification on this approach [104] linked the current instance instead with the antecedent which is classified as true and has the maximum likelihood, i.e., the best antecedent. Though this method had an overhead of processing all possible antecedent before conclusively deciding on one, the state-of-the-art model [79] also uses a version of this clustering albeit by restricting the search-space of the antecedent. Another kind of clustering deployed to generate the NP partition was the correlational clustering algorithm [88]. This algorithm measured the degree of inconsistency incurred by including a node in a partition and making repairs. This clustering type was different from the ones discussed earlier as the assignment to the partition was not only dependent on the distance measure with one node but on a distance measurement between all the nodes in a partition. For example, this clustering type avoided assigning the reference *she* to a cluster containing *Mr. Clinton* and *Clinton*. Though the classifier could have predicted that *Clinton* is an antecedent of *she* this link was avoided by using correlational clustering. Another variant of clustering algorithms used graph-partitioning. The nodes of the graph represented the mentions and the edge weights represented the likelihood of assignment of the pairs. Bell trees [82] were also used for creating an NP partition. In a Bell Tree, the root node was the initial state of the process which consisted of a partial entity containing the first mention of the document. The second mention was added in the next step by either linking to the existing mention or starting a new mention. The second layer of nodes was created to represent possible outcomes and subsequent mentions are added to the tree in a similar manner. The process was mention synchronous and each layer of the tree nodes was created by adding one mention at a time.

Another direction of research in mention pair models attempted at combining the phases of classification and effective partitioning using Integer Linear Programming [36,46]. As posited by Finkel and Manning [46] this task was suitable for integer linear programming (ILP) as CR required to take into consideration the likelihood of two mentions being co-referent during two phases: pair-wise classification and final cluster assignment phase. The ILP models first trained a classifier over pairs of mentions and encode constraints on top of probability outputs from pairwise classifiers to extract the most probable legal entity assignments. The difference between two ILP models mentioned earlier was that the former does not enforce transitivity while the latter encodes the transitivity constraint while making decisions. However, the ILP systems had a disadvantage that ILP is an NP-hard problem and this could create issues when the length of the document decreased. Another recently proposed model which eliminated the classification phase entirely was the algorithm by Fernandes et al. [44]. Their model had only two phases of mention detection and clustering. The training instances were a set of mentions x in the document and the correct co-referent cluster y. The training objective was a function of the cluster features (lexical, semantic, syntactic, etc.). This algorithm achieved an official CoNLL score of 58.69 and was one of the best performing systems in closed track of CoNLL 2012 shared-task.

In spite of being a widely used model for CR, there were some fundamental deficits with the mention-pair model. The first one was the constraint of transitivity which was enforced but did not always hold true. This meant that if an mention A referred to mention B and mention B referred to mention C, it was not always true that A co-referred with C, e.g., consider the case when *she* is predicted antecedent of *Obama* and *Obama* is predicted antecedent of *he*, but since *he* is not co-referent with *she* by violation of gender constraint, transitivity condition should not be enforced here. This flaw was mainly because the decisions made earlier by the coreference classifier were not exploited to correct future decisions. The information from only two NP's here *Obama* and *he* did not suffice to make an informed decision that they are co-referent, as the pronoun here was semantically empty. In addition, the NP *Obama* was itself ambiguous and could not be assigned any semantic feature like gender. Another disadvantage of the mention-pair model was that it only determined how good an antecedent was with respect to the anaphoric NP and not how good it was with respect to other antecedents available. The entity-mention models and the mention-ranking models were proposed with the aim of overcoming the disadvantages of the mention-pair models.

The entity mention model for CR focuses on a single underlying entity of each referent in discourse. This genre of algorithms was motivated by the fact that instead of making coreference decisions independently for every mention-antecedent pair it was necessary to exploit the past coreference decisions to inform the future ones. The entity mention model aimed at tackling this “expresiveness” issue [101] with the mention-pair model by attempting to classify whether an NP was co-referent with a preceding partially formed cluster instead of an antecedent. Thus, the training instances for the classifier were modified to a pair of NP N and cluster C and a label depicting whether the assignment of the NP to the partial cluster was positive or negative. Instances were represented as cluster-level features instead of pair wise features. The cluster-level features, e.g., gender and number, were defined over subsets of clusters using the “ANY”, “ALL”, “MOST”, etc. predicates. Entity mention model was evaluated by many researchers [82,161]. The former evaluated the entity-mention model in comparison to mention pair model on the ACE datasets using the decision tree classifier and inductive logic programming. The results for the entity-mention model as compared to the mention-pair model showed a slight decrease in performance using C4.5 classifier and a marginal increase in performance using inductive logic programming. The “ANY” constraint to generate cluster-level features was also encoded by the Bell Tree algorithm [82]. However, even in this case the performance of the entity mention model was not at par with the mention-pair model. The major reason for this was that it was extremely difficult to define cluster-level features for the entity-mention model. Most of the referents did not contribute anything useful to the cluster features because they were semantically empty (e.g., pronouns). Another model which attempted using features defined over clusters for CR was the first order probabilistic model by Culotta et al. [32]. Most recent models [26,28] also attempt at learning cluster-level features.

Mention-pair models faced an issue that they used a binary classifier to decide whether an antecedent was co-referent with the mention. The binary classifier could only provide a “YES” or “NO” result and failed to provide an intuition on how good one antecedent was compared to the other antecedent. The ranking models circumvented this flaw by ranking the mentions and choosing the best candidate antecedent. Ranking was considered to be a more natural way to predict the coreference links as it captured the competition between different antecedents. Some proposed models to realize this purpose were the tournament models and the twin candidate model by Yang et al. [159]. On a closer observation, the earlier rule-based approaches [65,73] also used constraints or sieves in a hierarchical manner starting with the most crucial ones to converge to the best antecedent. Hence, they too in principle ranked the antecedents using constraints which were ordered by their importance. A particularly prominent work which incorporated mention-ranking was the algorithm by Dennis and Baldridge [37] who replaced the classification function by a ranking loss. Another mention ranking model which used only surface features [42] and deployed a log-linear model for antecedent selection, outperformed the Stanford system [76] which was the winner of CoNLL 2011 shared task [117] by a margin of 3.5% and the IMS system [11] which was the then best model for CR by a margin of 1.9%.

Table 5
Mention pair variants.

Algorithm	NP Partitioning Algorithm	Learning Algorithm	Dataset	Performance metrics		
				Accuracy	MUC	B^3
[89]	Used Symmetricity and Transitivity to link mentions	Decision Tree C4.5	English Joint Venture Articles	86.5	-	-
			MUC-6	-	47.2	-
[134]	Closest first clustering	Decision Tree C5	MUC-6	-	62.6	-
			MUC-7	-	60.4	-
[104]	Best first clustering	RIPPER	MUC-6	-	70.4	-
		Decision Tree	MUC-7	-	63.4	-
[8]	Best first clustering	Averaged Perceptron Algorithm	ACE-Culotta test	-	75.8	80.8
[36]	Global inference with Integer Linear Programming	Maximum Entropy Model	ACE-BNEWS	-	69.2	-
			ACE-NPAPER	-	72.5	-
			ACE-NWIRE	-	67.5	-
[46]		Logistic Classifier	MUC-6	-	68.3	64.3
			ACE-NWIRE	-	61.1	73.1
			ACE-BNEWS	-	67.1	74.5
[88]	Graph Partitioning Algorithm	Conditional Random Fields over hidden Markov	MUC-6	-	73.42	-
[105]	Graph Partitioning	Maximum Entropy Model	MUC-6	-	89.63	-
[87]	Correlational Clustering	Conditional Random Fields over hidden markov models	MUC-6	-	91.59	-

In spite of its wide spread popularity, the mention rankers were still not able to effectively exploit past decisions to make current decisions. This motivated the "cluster ranking" algorithms. The cluster ranking approaches aimed at combining the best of the entity-mention models and the ranking models. Recent deep learning models [28] have also used a combination of mention ranker and cluster ranker for CR. Another issue with the mention-ranking model was that it did not differentiate between anaphoric and non-anaphoric NP's. The recent deep learning based mention ranking models [27,28,156,157] overcome this flaw by learning anaphoricity jointly with mention ranking. One of the earlier machine learning approaches which aimed at achieving this was the work of [122]. We provide a summary of the statistical and machine learning approaches in Table 5.

Until recently, the best performing model on the CoNLL 2012 shared task was an entity centric model [26]. Like other machine learning approaches, it also was feature rich. Defining features for mentions and especially for clusters is a very challenging task. Also, the extraction of the features is a time consuming task. This slowly started to change with the introduction of deep learning for NLP.

7.4. Deep learning models for coreference resolution

Since its inception, the aim of reference resolution research has been to reduce the dependency on hand-crafted features. With the introduction of deep learning in NLP, words could be represented as vectors conveying semantic dependencies [91,110]. This gave an impetus to approaches which deployed deep learning for CR [27,28,79,156,157].

The first non-linear mention ranking model [156] for CR aimed at learning different feature representations for anaphoricity detection and antecedent ranking by pre-training on these two individual subtasks. This approach addressed two major issues in CR: the first being the identification of non-anaphoric references in the text and the second was the complicated feature conjunction in linear models which was necessary because of the inability of simpler features to make a clear distinction between truly co-referent and non-coreferent mentions. This model handled the above issues by introducing a new neural network model which took only raw unconjoined features as inputs and attempted to learn intermediate representations.

The algorithm started with liberal mention extraction using the Berkeley Coreference resolution system [42] and sought to capture relevant aspects of the task better using representation learning. The authors proposed an extension to the original mention-ranking model using a neural network model, for which the scoring function is defined as:

$$s(x,y) \triangleq \begin{cases} u^T g\left(\begin{bmatrix} h_a(x) \\ h_p(x,y) \end{bmatrix}\right) + u_0 & \text{if } y \neq \epsilon \\ v^T h_a(x) + v_0 & \text{if } y = \epsilon \end{cases} \tag{25}$$

$$h_a(x) \triangleq tanh(W_a \theta_a(x) + b_a) \tag{26}$$

$$h_p(x,y) \triangleq tanh(W_p \theta_p(x,y) + b_p) \tag{27}$$

Hence, h_a and h_p represented the feature representations which were defined as non-linear functions on mention and mention-pair features θ_a and θ_b, respectively, and the function g's two settings were a linear function g_1 or a non-linear (tanh) function g_2, on the representations. The only raw features defined were θ_a and θ_p. According to the model, $C'(x)$ corresponded to the cluster the mention belongs to or ϵ if the mention was non-anaphoric. y_n^l corresponded to the highest scoring antecedent in cluster $C'(x)$ and was ϵ if x was non-anaphoric. The neural network was trained to minimize the slack rescaled latent-variable loss which the authors define as:

$$L(\theta) = \sum_{n=1}^{N} max_{\hat{y} \in \mathcal{Y}(x_n)} \Delta(x_n, \hat{y})(1 + s(x_n, \hat{y}) - s(x_n, y_n^l)) \tag{28}$$

Δ was defined as a mistake-specific cost function. The full set of parameters to be optimized was $W, u, v, W_a, W_p, b_a, b_p$. Δ could take on different values based on the type of errors possible in a CR task [42], i.e., false link(FL), false new(FN) and wrong link(WL), error types.

The subtask of anaphoricity detection aimed at identifying the anaphors amongst the extracted mentions. Generally the extracted mentions were non-anaphoric, thus this subtask served as an important step to filter out the mentions which needed further processing for antecedent ranking. The pre-trained parameters from this task were used for initializing weights of the antecedent ranking task. The antecedent ranking task was undertaken after filtering the non anaphoric mentions from the antecedent discovery process. The scoring procedure followed was similar to one discussed earlier.

The model was trained on two sets of BASIC [42] and modified BASIC + raw features. The baseline model used for anaphoricity prediction was an L1-regularized SVM using raw and conjoined features. The baseline model used for subtask two was the neural network based non-linear mention ranking model using the margin-based loss. The proposed neural network based model outperformed the two baseline models for both of the subtasks. The full model (g1 and g2) also achieved the best

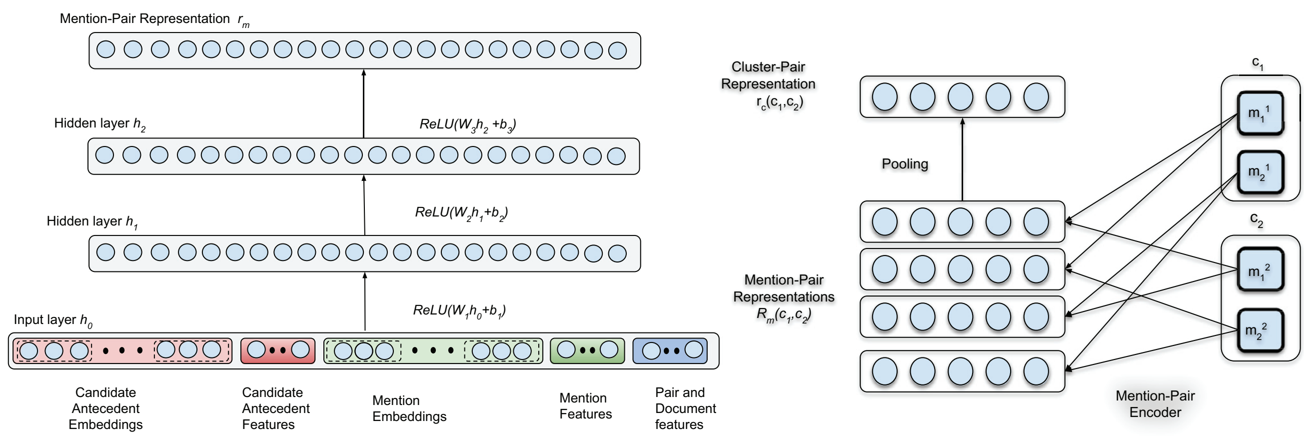

Fig. 4. The Mention-pair and the Cluster-pair encoder [28].

F_1 score with improvement of 1.5 points over the best reported models and 2 over the best mention ranking system(BCS). It outperformed all the state-of-the-art models (as of 2014) [11,42].

The first non-linear coreference model which proved that coreference task could benefit from modeling global features about entity clusters [157] augmented the neural network based mention-ranking model [156] by incorporating entity-level information produced by a recurrent neural network (RNN) running over the candidate antecedent-cluster. This model modified the scoring function of the antecedent ranking model by adding a global scoring term to it. The global score aimed to capture how compatible the current mention was with the partially formed cluster of the antecedent. The clusters were represented using separate weight sharing RNNs which sequentially consumed the mentions being assigned to each cluster. The idea was to capture the history of previous decisions along with the mention-antecedent compatibility. The Clark and Manning algorithm which was proposed roughly during the same time [28] instead defined a significantly different cluster ranking model to induce global information.

This approach was based on the idea of incorporating entity-level information, i.e., features defined over clusters of mention pairs. The architecture of this neural network consisted of mainly four sub-parts which were a the mention-pair encoder which passes features (described later) through a feed-forward neural network (FFNN) to produce distributed representations of mentions, a cluster-pair encoder which uses pooling over mention pairs to produce distributed representations of cluster pairs, a mention ranking model to mainly pre-train weights and obtain scores to be used further in cluster ranking and the cluster ranking module to score pairs of clusters by passing their representations through a single-layer neural network (Fig. 4).

The features used for the entire model were: the average of the embeddings of words in each mention, binned distance between the mentions, head word embedding, dependency parent, first word's, last word's and two preceding word's embedding and average of 5 preceding and succeeding words of the mention, the type of mention, position of mention, sub-mention, mention-length, document genre, string match, etc. These features were concatenated into an input vector and fed into a FFNN consisting of three fully-connected hidden rectified linear layers. The output of the last layer was the vector representation of the mention pair. The cluster pair encoder, given the two clusters of mentions c_i $=m^i_1, m^i_2, \ldots, m^i_{|ci|}$ and $c_j = m^j_1, m^j_2, \ldots, m^j_{|cj|}$, produces a distributed representation $r_c(c_i, c_j) \in R^{2d}$. This matrix was constructed by using max and average pooling over the mention-pair representations. Next, a mention-pair model was trained on the representations produced by the mention pair encoder which servers the purpose of pre-training weights for the cluster ranking task and to provide a measure for coreference decisions. This mention ranking model was trained on the slack rescaled objective [156] discussed earlier. The final stage was cluster ranking which used the pre-trained weights of the mention ranking model to obtain a score by feeding the cluster representations of the cluster encoder to a single-layered fully-connected neural network. The two available actions based on scores were merge (combine two clusters) and pass (no action). During inference, the highest-scoring (most probable) action was taken at each step. This ensemble of cluster ranking beat the earlier state-of-the-art approaches achieving an F1 score of 65.39 on the CoNLL English task and 63.66 on the Chinese task.

Another algorithm proposed by Clark and Manning [28] which complemented the earlier work by Clark and Manning [27] attempted at effectively replacing the heuristic loss functions which complicated the training, with the reinforce policy gradient algorithm and reward-rescale max-margin objective. This complementary approach exploited the importance of independent actions in mention ranking models. The independence of actions implied that the effect of each action on the final result was different thus making this scenario a suitable candidate for reinforcement learning. This model used neural mention ranking model [28] described earlier as the baseline and replaced the heuristic loss with reinforcement learning based loss functions. Reinforcement learning was utilized here to provide a feedback on different set of actions and linkages performed by the mention ranking models. Thus, the model could optimize its actions in such a way that the actions were performed to maximize the reward (called the reward rescaling algorithm). The reward rescaling algorithm achieves an average F_1 score of 65.73 and 63.4 on the $CEAF_{\phi 4}$ and B^3 metric respectively on the CoNLL 2012 English Test Data, thus outperforming the earlier systems. This algorithm was novel because it avoided making costly mistakes in antecedent resolution which could penalize the recall value. On the other hand, unimportant mistakes are not penalized as heavily. The approach is novel as it was the first one to apply reinforcement learning to CR. The most challenging task in this algorithm was the assignment of reward costs which could be corpus-specific.

The state-of-the-art model is an end-to-end CR system which outperformed the previous approaches in spite of being dependent on minimal features. This end-to-end neural model [79] is jointly modeled mention detection and CR. This model began with the construction of high-dimensional word embeddings to represent the words of an annotated document. The word embeddings used were a concatenation of Glove, Turian and character embeddings. The character embeddings were learnt using a character-level convolutional neural network (CNN) of three different window sizes. The vectorized sentences of the doc-

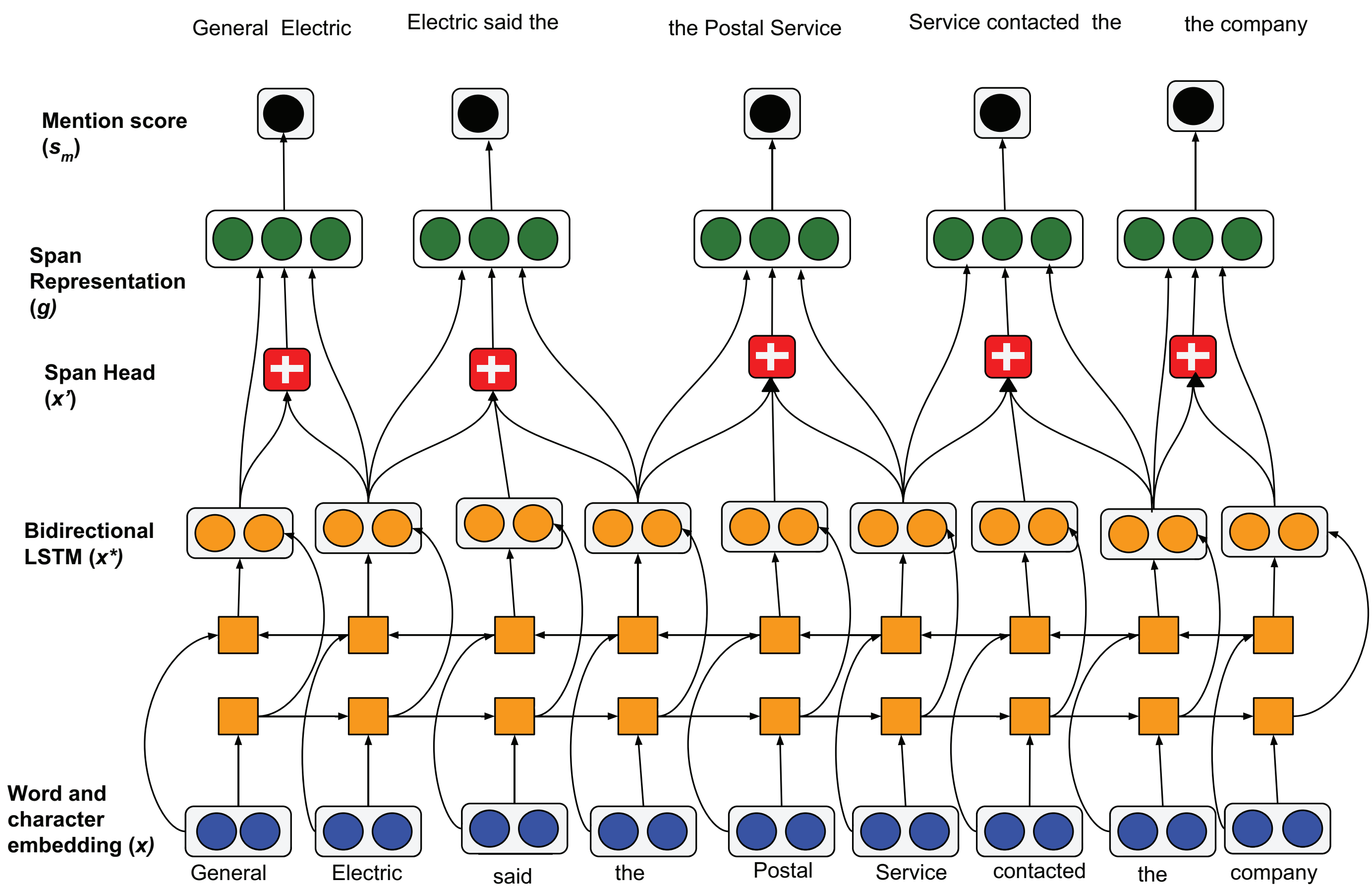

Fig. 5. Bi-LSTM to encode sentences and mention scoring [79].

ument were fed into a bidirectional long short-term memory (LSTM) network to learn effective word representations (Fig. 5). Next, all the possible mentions in a document were extracted and represented as a one dimensional vector. This mention representation was a conjugation of the start word embedding, head word embedding, end word embedding and some other mention-level features. The head word embedding was learnt using attention mechanism over the entire mention span. The mention representation g_i was defined as:

$$g_i = [x^*_{START(i)}, x^*_{END(i)}, x'_i, \phi(i)] \tag{29}$$

where x'_i represented an attention-weighted sum of the word vectors in span i and $x^*_{START(i)}$ and $x^*_{END(i)}$ are the span boundaries. The approach pruned candidates greedily for training and evaluation and considered only spans of maximum width ten. The mentions were scored using a FFNN and only a fraction of the top scoring spans were preserved further for CR.

These top scoring mentions served as input to the CR model (Fig. 6). The preceding 250 mentions were considered as the candidate antecedents. The scores of the mention-antecedent pairs were computed using the equation below. The mention-antecedent pair representation was a concatenation of individual mention representations g_i and g_j, the similarity between the two mentions $g_i \circ g_j$ and pairwise features $\phi(i, j)$ representing speaker and distance features. The final scoring function optimized is a sum of the of the two individual mention scores of the candidate mentions and the mention-antecedent pair score represented by the equation below.

$$s_a(i, j) = w_a \cdot FFNN_a([g_i, g_j, g_i \circ g_j, \phi(i, j)]) \tag{30}$$

The optimization function used for the model was the marginal log-likelihood of all correct antecedents on basis of the gold-clustering.

$$\log \prod_{i=1}^{N} \sum_{y' \in \mathcal{Y}\hat{i}GOLD(i)} P(y') \tag{31}$$

During inference the best scoring antecedent was chosen as the most probable antecedent and coreference chains were formed using the property of transitivity.

The authors report the ensembling experiments using five models with different initializations and prune the spans here using average of the mention scores over each model. The proposed approach was extensively evaluated for precision, recall and F1 on MUC, B^3 and CEAF metrics. The authors also provide a quantitative and qualitative analysis of the model for better interpretability.

Challenging aspect of this model is that its computational time is high and large number of trainable parameters need to be stored. This model used a very large deep neural network and, hence, is very difficult to maintain. This creates a challenge for deploying this system as an easy to use off-the-shelf system.

Deep learning CR systems [27,28,79] represent words using vectors which are known depict semantic relationships between words [91,110]. These models, hence, use less features than the machine learning models. These systems also implicitly capture the dependencies between mentions particularly using RNN and its adaptations like LSTMs and gated recurrent units (GRUs). One disadvantage of these systems is that they are difficult to maintain and often require some amount of genre- or domain-specific adaptation before use. Amongst the deep learning based CR algorithms discussed earlier, we observe that the dependency on features decreased over time. This was mainly because of

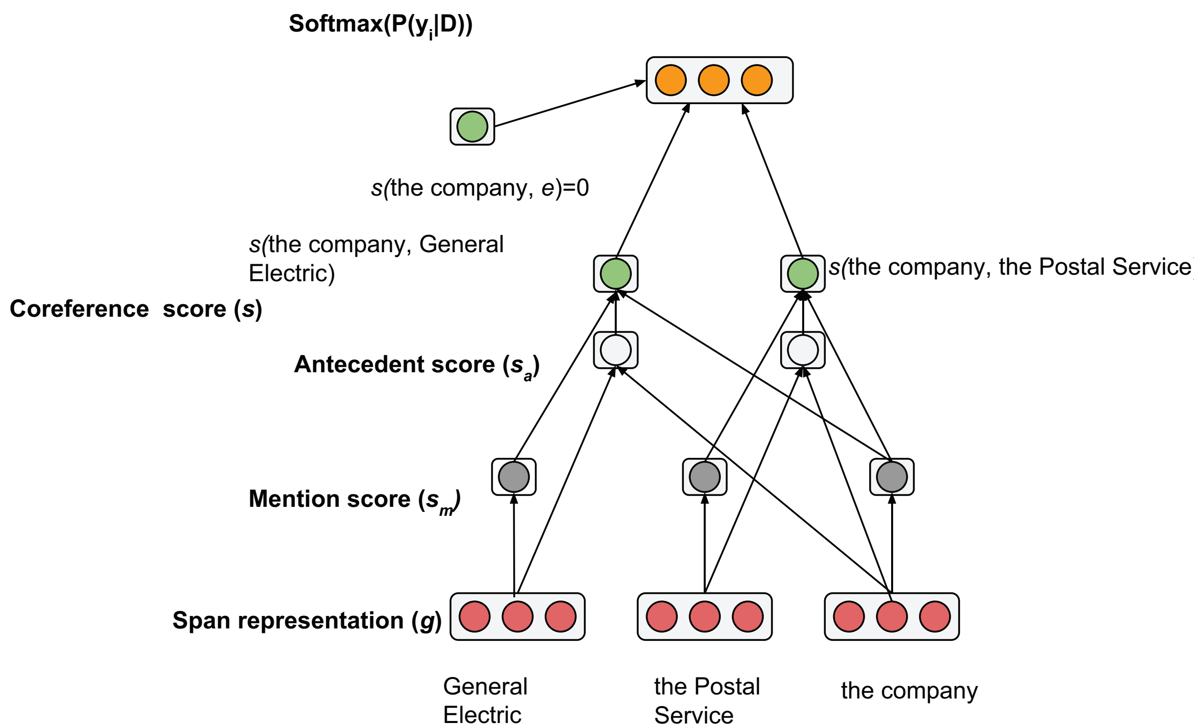

Fig. 6. Antecedent Scoring [79].

the pre-trained word embeddings which captured some amount of semantic similarity between the words. Unlike the Stanford deep coref system [27,28], the state-of-the-art system used minimal mention-level and mention-antecedent pair features. This system also did not use any external mention-extractor and proceeds by extracting all possible spans up to a fixed width, greedily. Another advantage of this system was that it did not use any heuristic loss function unlike the other deep learning models [27,28,156,157] and still managed to beat the earlier models with a very simple log-likelihood loss function. The previous models used heuristic loss functions which were dependent on a mistake-specific cost function whose values were set after exhaustive experiments. Though this system is difficult to maintain mainly because of its high-dimensionality, it is a strong evidence of the effectiveness of LSTMs and their ability to capture long term dependencies. Another possible disadvantage of this model is that it is still basically a mention ranking model and chooses the highest scoring antecedent without using any cluster-level information. As posited by many earlier deep learning works which used cluster-level information [28,157] this information is necessary to avoid linking of incompatible mentions to partially formed coreference chains. In spite of some of the disadvantages of the deep CR systems, future strides in CR can only be achieved by either defining better features to be learnt or by introducing better deep learning architectures for CR. We have summarized the deep learning models for coreference resolution in Table 6.

7.5. Open-source tools

In this section, we provide a summary of the open-source tools available for the task, most of which are based on the algorithms described earlier. Providing an open-source implementation of algorithms, helps researchers think about possible improvements from peer suggestions and also allows researchers mainly interested in its application to pick an off-the-shelf model. These tools may deploy a specific approach like [97] and or others like Reconcile [137] could be a combination of many research methodologies.

The domain dependent GuiTAR tool [113] aimed at making an open-source tool available for researchers mainly interested in applying AR to upstream NLP applications. Stanford coref toolkit provides 3 models which were pioneered by the Stanford NLP group. These three algorithms are Deterministic [75,121,126], Statistical [26] and Neural [27,28]. The multilingual BART [147] tool is one of the few highly modular toolkit for CR that supports the statistical approaches. It relies on a maximum entropy model for classification of mention-pairs. BART proceeds by converting input document into a set of linguistic layers represented by separate XML layers. They are used to extract mentions, assign syntactic properties and define pairwise features for the mention. A decoder generates training examples through sample selection and learns pairwise classifier. The encoder generates testing examples through sample selections and partitions them based on trained coreference chains. This toolkit aimed at combining the best state-of-the-art models into a modular toolkit which has been widely used for broader applications of AR. ARKref [107] is a tool for NP (Noun Phrase) CR that is based on system described by Haghighi and Klein [57]. ARKref is deterministic, rule-based that uses syntactic information from a constituency parser, semantic information from an entity recognition component to constraint the set of possible antecedent candidate that could be referred by a given mention. It was trained and tested on CoNLL shared task [116]. The Reconcile System [137] solved a problem of comparison of various CR algorithms. This problem mainly arises due to high cost of implementing a complete end to end CR system, thus giving way to inconsistent and often unrealistic evaluation scenarios. Reconcile is an infrastructure for development of learning based NP CR system Reconcile can be considered a combination of rapid creation of CR systems, easy implementation of new feature sets and approaches to CR and empirical evaluation of CR across a variety of benchmark datasets and scoring metrics. It aims to address one of the issues in AR [94], which is the huge disparity in the evaluation standards used. It further makes an attempts to reduce the labelling standards disparity too. We have summarized the open source tools available for the tasks in Table 7.

Table 6
Deep learning based coreference resolution.

Algorithm	Neural Network architecture(s) used	Pre-trained Word Embeddings Used	Cluster-level features used (Y/N)	Loss function used for Mention Ranking	External Tools Used
[156]	FFNN	-	No	Heuristic Regularized Slack rescaled latent variable Loss	Berkeley Coreference System for mention extraction and Stanford Coref System's Rules for animacy feature
[157]	FFNN and RNN	-	Yes	Heuristic Slack Rescaled Margin objective	Berkeley Coreference System for mention extraction and Stanford deterministic system animacy rules
[28]	FFNN	English:50d word2vec Chinese:Polyglot 64d	Yes	Heuristic Slack Rescaled Max-margin objective	Stanford Deterministic Coref System rules to extract mentions
[27]	FFNN	English:50d word2vec Chinese:Polyglot 64d	Yes	Heuristic Max-margin objective, REINFORCE policy gradient, Reward Rescaling Algorithm	Stanford Deterministic Coref System rules to extract mentions
[79]	FFNN+Bi-LSTM + CNN + Neural Attention	Glove300d+turian 50d	No	Marginal Log-likelihood	-

Table 7
Off-the-shelf reference resolution systems.

Toolkit	Algorithm	evelopment and Evaluation Corpus	Languages	Type
GuiTAR [113]	General toolkit which incorporates many algorithms like [97], [149], etc. and can be extented to include other algorithms	GNOME corpus	English	Hybrid: Rule-based+ Machine Learning
BART [147]	Primarily [134] and some other machine learning approaches to CR	ACE-2 corpora	English	Machine Learning
PARKref [107]	PHaghini and Klein Model [57]	ACE2004-ROTH-DEV ACE2004-CULOTTA -TEST	German, English and Italian	Rule-based
Reconcile [137]	Abstracts the basic architecture of most contemporary supervised learning- based coreference resolution systems e.g., [8,104,134]	2 MUC datasets (MUC-6 and MUC-7) 4 ACE datasets	English	Supervised Machine Learning Classifiers
Stanford CoreNLP deterministic, rule-based system	[75]	CoNLL 2011 (OntoNotes Corpus), ACE2004-Culotta- Test, ACE 2004-nwire, MUC6-Test	Chinese English	Rule-based
Stanford Core NLP Statistical System	[26]	CoNLL 2012	English	Statistical
Stanford CoreNLP Neural Coreference Resolution	[27,28]	CoNLL 2012	English	Deep Neural Network

8. Reference resolution research progress on different datasets

In previous sections, we have discussed several types of reference resolution algorithms. In this section, we aim at providing an overview of the research progress made in the field of reference resolution over the past few years. Here, we analyze the progress made on mainly three important publicly available datasets: MUC, ACE and CoNLL shared task corpus (OntoNotes 5.0).

The MUC datasets were the first annotated corpora to be publicly available for CR. Though these datasets were small they have been widely used for evaluation and training. The first system to be evaluated on the MUC dataset was Soon's mention-pair model [134]. This was followed by Ng and Cardie's series of improvements on the dataset [103,104]. Followed by these were some mention ranking models which were also attempted on the MUC datasets. One of them was Yang's twin candidate model [159] which aimed at capturing competition between the antecedents. Conditional random fields (CRFs) [88] have also been trained on this dataset to directly model global coreference chains. In addition, some other approaches like Integer Linear Programming [46] and non-parametric Bayesian models [56] have also been attempted. The MUC-6 and 7 datasets in spite of being widely popular were quite small in size thus making training on a small corpus very hard. This also meant that some types of references were ignored. The evaluation standards were also not very well defined, hence making comparison of different algorithms a challenge. The ACE and CoNLL datasets aimed at overcoming these disadvantages by providing a much larger training corpus.

When coming to the ACE datasets, we observe a huge disparity in the evaluation standards, train-test splits and metrics used. This was mainly because the test sets of the dataset were not publicly released and, hence, were unavailable to non-participants. This made comparative evaluation with research methodologies which did not participate in this task difficult. Many researchers were hence forced to define their own train-test split [8,32]. In addition, the ACE datasets were also released in iterations and phases from 2002 to 2005, thus algorithms tested on newer releases could not be directly compared with the earlier approaches. Multiple algorithms were evaluated on different versions of the ACE datasets like the mention pair models [8], mention ranking models [37,122] and joint inference models [58,114]. Some rule-based approaches [76] were also tested on the ACE datasets mainly with the aim of comparison with past research methodologies, which were not evaluated on the newly introduced CoNLL shared task.

The best performing rule-based systems on the current version of the CoNLL shared task is the multi-sieve based Stanford deterministic system [75]. Most of the early systems which outperformed the rule-based system were machine learning based. There have been multiple variants of the mention-pair models which used structured perceptron models for CR on the CoNLL 2012 dataset [21,42,44]. This was followed

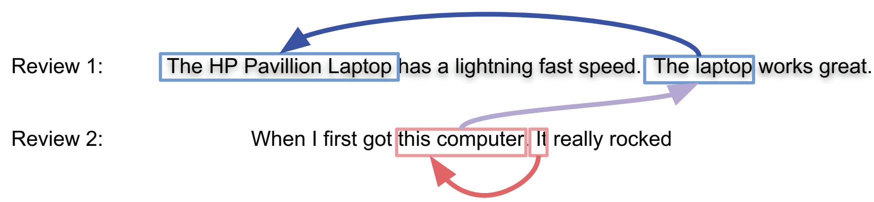

Fig. 7. Global Reference Resolution (Neural Coref Api).

by a model which jointly modeled Coreference, Named Entity Recognition and Entity Linking using a structured CRF. Models which used cluster-level features were very well known in CR by then and some models [83] also used average across all pairs in clusters involved to incorporate cluster-level features. The entity-centric model [26] which achieved a large margin of improvement on the earlier systems proposed a novel way of defining cluster-level features. It combined information from involved mention pairs in variety of ways with higher order features produced from scores of the mention pair models. As observed in Table 8, since 2015 the best performing systems on CoNLL 2012 datasets have been deep learning systems. These used neural networks to individually model different subtasks like antecedent ranking and cluster ranking or to jointly model mention prediction and CR. Though the most common use of deep neural networks in CR has been for scoring mention-pairs and clusters, some methods [157] also used RNN's to sequentially store cluster states, with the aim of modeling cluster-level information. The best performing system on the CR task is a very deep end to end system which is uses a combination of LSTM, CNN, FFNN and neural attention. Since deep learning systems are typically hard to maintain some recent systems have also proposed a hybrid of rule-based and machine learning systems [78]. Though this system does not perform at par with the deep learning system, it is easy to maintain and use and even outperforms some of the machine learning systems. Overall, the deep learning trend in CR looks very exciting and future progress could be expected by incorporating better features and using more sophisticated neural network architectures. As stated by many researchers [42,79], this could be modeled by developing an intuitive way to incorporate differences between entailment, equivalence and alteration phenomenon.

As observed in Table 8, until about a few years ago we observe that the CR datasets and mainly the evaluation metrics were not standardized. This made comparison of algorithms very difficult. The early corpora like MUC and ACE did not release very strict evaluation guidelines for the task. Also, there were multiple releases, only few of which were publicly available. The test datasets of the ACE corpora were initially not available to non-participants which also created issues with the comparison of algorithms. Hence, most authors often defined train and test splits of their own on the datasets [8,32]. Though future approaches tried to stick to the earlier train-test splits for comparative evaluation [57], it was difficult as often the datasets needed for comparison were not freely available. Another issue was with the very definition of the Coreference Task. Some approaches [82] which reported highest accuracy on the ACE and MUC corpus could not be compared with others because they reported performance on true labelled mentions instead of system predicted mentions. This was different from other approaches which jointly modeled the tasks of mention prediction and CR. This, however, changed with the introduction of CoNLL 2012 shared task [116] which defined strict evaluation guidelines for CR. After this, CR research gained momentum and has seen more consistent progress and clearer evaluation standards.

9. Reference resolution in sentiment analysis

Being one of the core component of natural language understanding, CR has many potential downstream applications in NLP like machine translation [6,61,154], paraphrase detection [129,130], summarization [10,158], question answering [148,155,162], and sentiment analysis [17,146]. Sentiment analysis, in particular, has recently raised growing interest both within the scientific community (leading to many exciting open challenges) as well as in the business world (due to the remarkable benefits to be had from marketing and financial market prediction). Though AR is said to be one of the most commonly faced challenges in sentiment analysis, we observe that there is scarcity of research work targeting the question of how could CR be effectively incorporated into a sentiment analysis system (and which specific issues it could solve). Here, we provide potential scenarios which necessitate CR. We also discuss some of the prominent approaches at the intersection of these two fields. We believe that state-of-the-art approaches for sentiment analysis could highly benefit from additional reference resolution information. The goal of the section is to encourage future evaluations and verifications in this direction.

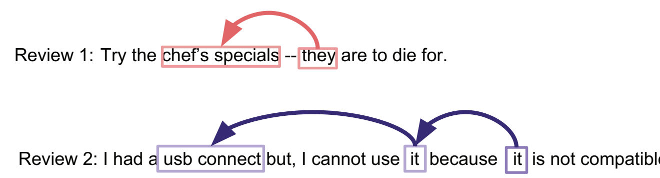

Fig. 8. Fine-grained aspect resolution (Neural Coref Api).

While analyzing the use of AR in sentiment analysis, we came across two main scenarios where AR can prove beneficial, the first being global reference resolution. An often observed phenomenon in product or service reviews is that they are often centered around one particular entity. Hence, most reviewers do not explicitly specify the entity that corresponds to their opinion target. Thus, cross-review information can be exploited effectively to resolve the pronominal references to the global entity. An example of this (taken from the SemEval aspect-based sentiment analysis dataset) is depicted in Fig. 7. In the example, multiple reviews could be used to chain the references to a global entity (i.e., HP Pavillion Laptop). Global reference resolution can aid the process of extracting the sentiment associated with the general entity.

Another possible use of AR in sentiment analysis is mainly for fine-grained aspect-based sentiment analysis [84]. AR can help infer multiple pronominal references to a particular aspect of the product. This, in turn, can help extract the opinion associated with that particular aspect. An example of this can be seen in Fig. 8, where the resolution of the pronouns related to the aspects (as depicted by the links between them) can aid in the extraction of fine-grained (aspect-specific) polarity. These two images were the resolved references returned by the hugging face api, which deploys the Stanford Deep Coref System [27].

Now that we have established the importance of AR for sentiment analysis, we provide an overview of the approaches which have worked at the intersection of these two fields. The importance of AR in sentiment analysis has been delineated in many significant research works which consider sentiment analysis as a suitcase research problem [16]. AR and CR enable sentiment analysis to break beyond the sentence-level opinion mining task. A recent approach which targets this [74] addresses the problem of aspect extraction, a crucial sub-task of aspect-based sentiment analysis. This approach leverages an ontological resolution engine which discovers the “implied by” and “mentioned by” relations in the aspect-based sentiment ontology. Another paper [106] aimed

Table 8
Dataset-wise comparison of baselines.

Dataset	Release	Algorithm	Scoring metrics F1 values				Algorithm Type
			MUC	B^3	CEAFe	CoNLL	
CoNLL shared task (OntoNotes 5.0)	CoNLL 2011	[76]	61.51	63.27	45.17	56.65	Rule-based
		[11]	64.71	64.73	45.35	58.26	Machine Learning
		[42]	66.43	66.16	47.79	60.13	
	CoNLL 2012	[75]	63.72	52.08	48.65	54.82	Rule-based
		[21]	69.48	57.44	53.07	60.00	Probabilistic
		[44]	70.51	57.58	53.86	60.65	Machine Learning
		[42]	70.51	58.33	55.36	61.40	
		[83]	72.84	57.94	53.91	61.56	
		[11]	70.72	58.58	55.61	61.63	
		[43]	71.24	58.71	55.18	61.71	
		[26]	72.59	60.44	56.02	63.02	
		[78]	72.37	60.46	56.76	63.20	
							Hybrid=ML+Rules
		[156]	72.6	60.52	57.05	63.39	Deep Learning
		[157]	73.42	61.50	57.7	64.21	
		[28]	74.06	62.86	58.96	65.29	
		[27]	74.56	63.40	59.23	65.73	
		[79]	77.20	66.60	62.60	68.80	
Automatic Content Extraction	ACE 2004 Culotta Test	[137]	62.0	76.5	-	-	Machine Learning
		[57] (true)	79.60	79.00	-	-	Rule-based
		[57] (system)	64.4	73.2	-	-	Rule-based
		[58]	67.0	77.0	-	-	Machine Learning
		[32]	-	79.30	-	-	Probabilistic
		[8]	75.80	80.80	-	-	Machine Learning
		[75]	75.90	81.00	-	-	Rule-based
	ACE 2004 -training datasets BNEWS,NWIRE	[56]	62.3,64.2	-	-	-	Rule-based
		[46]	67.1,61.1	74.5,73.1	-	-	ML+ Integer Linear Programming
		[114]	70.90,67.3	-	-	-	Machine Learning
		[57]	76.50,-	76.90,-	-	-	Rule-based
		[75]	79.60,-	80.20,-	-	-	Rule-based
	ACE -Phase 2 Test sets BNEWS,NWIRE,NPAPER	[100]	64.9,54.7,69.3	65.6,66.4,66.4	-	-	Machine Learning
		[38]	69.2,67.5,72.5	-	-	-	
	-	[114]	67.4,67.4,70.4	67.7,71.6,68.2	-	-	
	ACE 2005 Stoyanov Test	[137]	67.4	73.7	-	-	
		[57]	65.2	71.8	-	-	Rule-based
		[58]	68.1	75.1	-	-	Machine Learning
	ACE 2005 Rahman and Ng	[122]	69.3	61.4	-	-	
		[57]	67.0	60.6	-	-	Rule-based
		[58]	71.6	62.7	-	-	Machine Learning
Message Understanding Conference	MUC 6	[134]	62.6	-	-	-	Machine Learning
		[103]	69.5	-	-	-	
		[104]	70.4	-	-	-	
		[160]	71.3	-	-	-	
		[88]	73.4	-	-	-	
		[56]	63.9	-	-	-	
		[46]	68.30	64.30	-	-	ML+ Integer Linear Programming
		[137]	68.5	70.88	-	-	Machine Learning
		[114]	79.20	-	-	-	
		[57]	81.90	75.0	-	-	Rule-based
		[75]	78.40	74.40	-	-	
	MUC 7	[134]	60.4	-	-	-	Machine Learning
		[104]	63.4	-	-	-	
		[160]	60.2	-	-	-	
		[137]	62.8	65.86	-	-	

at investigating whether a performance boost is obtained on taking coreference information into account in sentiment analysis. Take, for example, the sentence "*The canon G3 power shot has impressed me. This camera combines amazing picture quality with ease of use*". For a human annotator, it is easy to understand that the term camera here co-refers with canon G3 power shot. However, this task is a major challenge faced by most algorithms. The sentiment analysis algorithm introduced here is proximity-based for focused sentiment identification. If first calculates the anchor-level sentiment by considering a sentiment window of 8 tokens before and after a phrase using distance weighting approach. The anchor weighted scores are aggregated and sentiment phrases are created. Finally, the co-referring entities are identified and the algorithm is evaluated over an opinionated corpus. The percentage improvement obtained over baseline CR modules is on an average 10% and varies over different datasets used for evaluation.

Another algorithm [67] aimed at tackling the issue of extracting opinion targets expressed by anaphoric pronouns. Opinion word and target pair extraction can benefit from AR to a great extent. The algorithm presented by Zhuang et al. [166] is used as a baseline for the experiment using opinion target and opinion word extraction. A modified version of CogNIAC [4] is used for CR. The best configuration of this algorithm reaches approximately 50% of the improvements which are theoretically possible with perfect AR.

Another recent interesting work [39] posits that object and attribute coreference is important because, without solving it, a great deal of opinion information will be lost and opinions may be assigned to wrong entities. The paper elicits the importance of this issue with an example: "*I bought the cannon S500 camera yesterday. It looked beautiful. I took a few photos last night. They were amazing*". Here, the last two sentences express opinions but it is difficult to specify the target at which the opinion is aimed. Target extraction becomes meaningless if the association between the target and the opinion word is not captured appropriately or is obscure due to co-referent phrases. The paper describes two basic entities object and attribute, e.g., camera (object) and picture quality (attribute). The pairwise learning approach adopted is a supervised model based on [134] CR feature model and the annotation is performed as per MUC-7 standards. The datasets used are blog conversations on products of multiple categories like dvd, cars, tv, lcd, etc. The algorithm first pre-processes the text and then constructs features in a way similar to [134] with addition of some other features like sentiment consistency, comparative sentences and entity-opinion word pair association. A decision tree is trained on these features and the algorithm is tested on an unannotated dataset.

As observed from the research methodologies discussed earlier, the merging of CR systems with sentiment analysis systems is a challenging task. This is further accentuated by the fact that current CR systems are themselves far from perfect and resolving references before sentiment analysis could in fact prove detrimental to polarity detection if not incorporated correctly. Future research methodologies in this area should focus on more exhaustive evaluations on standard sentiment analysis datasets and on jointly solving CR and polarity classification via multi-task learning, as it has been investigated for other sub-tasks of sentiment analysis, e.g., sarcasm detection [85].

10. Reference resolution: Issues and controversies

In this section, we discuss major issues and controversies spanning the area of reference resolution. Upon a thorough investigation of past research, we individuated three main areas of debate in this field: the evaluation metrics used, the scope of the datasets used and the idea of commonsense knowledge induction for reference resolution. We provide an overview of these issues and the progress made in addressing them.

The issues with the evaluation metrics to be used for CR have been delineated by many prominent researchers [94,101]. We progressed from using simple metrics like Hobb's algorithm to developing MUC [150], B^3 [3] and CEAF [82]. In spite of the progress made over recent years, however, the main evaluation method currently used for CR is still simply the average of those three metrics, which still present several issues [116]. Recently, some researchers proposed new metrics to circumvent the issues faced by the earlier ones, e.g., modifications of existing metrics [13] and a new LEA metric [98]. We encourage researchers to evaluate their models on these recently proposed metrics in addition to the earlier standard metrics.

Another area pertaining to CR research is whether the standard datasets for the task address different types of references that exist in natural language. As discussed earlier, the fields of CR and AR span many different types of references. Some of these references are rare and some types are not labelled by current CR datasets [165]. This has led to the proliferation of research targeting specific types of references like multi-antecedent references [145], Abstract Anaphora [86] and One Anaphora [50]. To avoid confusion in the community, we suggest that new datasets clearly specify the types of references they are addressing and the ones they are not. We also encourage future CR models to carry out cross-domain evaluations on other datasets which are also annotated in CoNLL format like the Character Identification dataset [98]. This will aid in the process of identifying the types of references that still pose a challenge for state-of-the-art CR algorithms.

Since early stages, it has been known that some types of references are extremely hard to resolve for a machine mainly because they require some amount of external world knowledge. Though the usefulness of world knowledge for a coreference system has been known since the late nineties, early mention pair models [104,134,160] did not incorporate any form of world knowledge into the system. As knowledge resources became less noisy and exhaustive, some CR researchers started deploying world knowledge to CR. The two main questions to be answered were whether world knowledge offered complementary benefits and whether the noisy nature of world knowledge would affect the performance of the model negatively. Several researchers have deployed world knowledge in the form of web-based encyclopaedias [144], unannotated data [35], coreference annotated data [8], and knowledge bases like YAGO, Framenet [123] and Wordnet [42]. World knowledge was mainly incorporated as features into the mention pair models and cluster ranking models. These features were often defined over NPs and verbs. Some initial algorithms reported an increase in performance up to 4.8% for inducing world-knowledge features from YAGO and FrameNet. While some others [42] reported only minor performance gains using world knowledge on system mentions extracted by CR systems. Instead of representing commonsense knowledge as features, some models used predicates to encode commonsense relations [109]. They evaluated their model on hard CR problems that fit the definition of the Winograd Schema Challenge. As posited by Durret and Klein [42], the task of modeling complex linguistic constraints into a coreference system remains an uphill battle.

11. Conclusion

In this survey paper, we presented an exhaustive overview of the field of coreference resolution and the closely related field of anaphora resolution, which form a core component of NLP research. We put forth a detailed account of the types of references and the important constraints necessary for resolution with the aim of establishing the broad scope of the task. We also clarified the boundaries between the tasks of coreference resolution and anaphora resolution to enable more focused research progress in the future. In addition, we compared predominantly used evaluation metrics. We observed that, although there are multiple datasets available, some state-of-the-art methods have not been evaluated on them. With the spirit of encouraging more exhaustive evaluations, we also provided an account of the datasets released for the task.

Reference resolution research has seen a shift from rule-based methods to statistical methods. To this end, we provided an analysis of the types of algorithms used with special focus on recent deep learning methods. We concluded the survey by taking a closer look at one popular NLP task that could highly benefit from reference resolution: sentiment analysis. As the research in the intersection of these two fields is scarce, we established a background for the inter-dependency between the two tasks. Finally, we listed outstanding issues in reference resolution research, thus laying a firm cornerstone for future researchers to build on.

Declaration of Competing Interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed.

Acknowledgments

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Projects #A18A2b0046 and #A19E2b0098).

References

[1] C. Aone, S.W. Bennett, Evaluating automated and manual acquisition of anaphora resolution strategies, in: Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 1995, pp. 122–129.
[2] J.D. Atlas, S.C. Levinson, It-clefts, Informativeness and logical form: radical pragmatics (revised standard version), in: Radical Pragmatics, Academic Press, 1981, pp. 1–62.
[3] A. Bagga, B. Baldwin, Algorithms for scoring coreference chains, in: The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference, 1, Granada, 1998, pp. 563–566.
[4] B. Baldwin, Cogniac: high precision coreference with limited knowledge and linguistic resources, in: Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts, Association for Computational Linguistics, 1997, pp. 38–45.
[5] R.T. Batista-Navarro, S. Ananiadou, Building a coreference-annotated corpus from the domain of biochemistry, in: Proceedings of BioNLP 2011 Workshop, Association for Computational Linguistics, 2011, pp. 83–91.
[6] R. Bawden, R. Sennrich, A. Birch, B. Haddow, Evaluating discourse phenomena in neural machine translation., 2017 arXiv:1711.00513.
[7] C.A. Bejan, S. Harabagiu, Unsupervised event coreference resolution, Comput. Linguist. 40 (2) (2014) 311–347.
[8] E. Bengtson, D. Roth, Understanding the value of features for coreference resolution, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 294–303.
[9] A.L. Berger, V.J.D. Pietra, S.A.D. Pietra, A maximum entropy approach to natural language processing, Comput. Linguist. 22 (1) (1996) 39–71.
[10] S. Bergler, R. Witte, M. Khalife, Z. Li, F. Rudzicz, Using knowledge-poor coreference resolution for text summarization, in: Proceedings of DUC, 3, 2003.
[11] A. Björkelund, R. Farkas, Data-driven multilingual coreference resolution using resolver stacking, in: Joint Conference on EMNLP and CoNLL-Shared Task, Association for Computational Linguistics, 2012, pp. 49–55.
[12] S.E. Brennan, M.W. Friedman, C.J. Pollard, A centering approach to pronouns, in: Proceedings of the 25th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 1987, pp. 155–162.
[13] J. Cai, M. Strube, Evaluation metrics for end-to-end coreference resolution systems, in: Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, 2010, pp. 28–36.
[14] E. Cambria, Affective computing and sentiment analysis, IEEE Intell. Syst. 31 (2) (2016) 102–107.
[15] E. Cambria, A. Hussain, C. Havasi, C. Eckl, Common sense computing: from the society of mind to digital intuition and beyond, in: J. Fierrez, J. Ortega, A. Esposito, A. Drygajlo, M. Faundez-Zanuy (Eds.), Biometric ID Management and Multimodal Communication, Lecture Notes in Computer Science, 5707, Springer, Berlin Heidelberg, 2009, pp. 252–259.
[16] E. Cambria, S. Poria, A. Gelbukh, M. Thelwall, Sentiment analysis is a big suitcase, IEEE Intell. Syst. 32 (6) (2017) 74–80.
[17] E. Cambria, S. Poria, D. Hazarika, K. Kwok, SenticNet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings, in: AAAI, 2018, pp. 1795–1802.
[18] J.G. Carbonell, R.D. Brown, Anaphora resolution: a multi-strategy approach, in: Proceedings of the 12th Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics, 1988, pp. 96–101.
[19] L. Carlson, D. Marcu, M.E. Okurowski, Building a discourse-tagged corpus in the framework of rhetorical structure theory, in: Current and New Directions in Discourse and Dialogue, Springer, 2003, pp. 85–112.
[20] L. Castagnola, Anaphora Resolution for Question Answering, Massachusetts Institute of Technology, 2002 Ph.D. thesis.
[21] K.-W. Chang, R. Samdani, D. Roth, A constrained latent variable model for coreference resolution, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 601–612.
[22] I. Chaturvedi, E. Cambria, R. Welsch, F. Herrera, Distinguishing between facts and opinions for sentiment analysis: survey and challenges, Inf. Fus. 44 (2018) 65–77.
[23] D. Chen, C. Manning, A fast and accurate dependency parser using neural networks, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 740–750.
[24] Y.-H. Chen, J.D. Choi, Character identification on multiparty conversation: identifying mentions of characters in tv shows, in: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2016, pp. 90–100.
[25] N.A. Chinchor, Overview of muc-7/met-2, Technical Report, SCIENCE APPLICATIONS INTERNATIONAL CORP SAN DIEGO CA, 1998.
[26] K. Clark, C.D. Manning, Entity-centric coreference resolution with model stacking, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 1, 2015, pp. 1405–1415.
[27] K. Clark, C.D. Manning, Deep reinforcement learning for mention-ranking coreference models., 2016 arXiv:1609.08667.
[28] K. Clark, C.D. Manning, Improving coreference resolution by learning entity-level distributed representations., 2016 arXiv:1606.01323.
[29] K.B. Cohen, H.L. Johnson, K. Verspoor, C. Roeder, L.E. Hunter, The structural and content aspects of abstracts versus bodies of full text journal articles are different, BMC Bioinformatic. 11 (1) (2010) 492.
[30] W.W. Cohen, Y. Singer, A simple, fast, and effective rule learner, AAAI/IAAI 99 (1999) 335–342.
[31] R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, in: Proceedings of the 25th International Conference on Machine Learning, ACM, 2008, pp. 160–167.
[32] A. Culotta, M. Wick, A. McCallum, 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, 2007, pp. 81–88.
[33] A. Cybulska, P. Vossen, Guidelines for ECB+ annotation of events and their coreference, Technical Report, Technical Report NWR-2014-1, VU University Amsterdam, 2014.
[34] W. Daelemans, J. Zavrel, K. Van Der Sloot, A. Van den Bosch, Timbl: tilburg memory-based learner, Tilburg Univer. (2004).
[35] H. Daumé III, D. Marcu, A large-scale exploration of effective global features for a joint entity detection and tracking model, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2005, pp. 97–104.
[36] P. Denis, J. Baldridge, 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, 2007, pp. 236–243.
[37] P. Denis, J. Baldridge, Specialized models and ranking for coreference resolution, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 660–669.
[38] P. Denis, J. Baldridge, Global joint models for coreference resolution and named entity classification, Procesamiento del Lenguaje Natural 42 (2009).
[39] X. Ding, B. Liu, Resolving object and attribute coreference in opinion mining, in: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, 2010, pp. 268–276.
[40] R.M. Dixon, Demonstratives: a cross-linguistic typology, Stud. Lang. Int. J. Sponsor. Foundat. "Foundations of Language" 27 (1) (2003) 61–112.
[41] G.R. Doddington, A. Mitchell, M.A. Przybocki, L.A. Ramshaw, S. Strassel, R.M. Weischedel, The automatic content extraction (ace) program-tasks, data, and evaluation., in: LREC, 2, 2004, p. 1.
[42] G. Durrett, D. Klein, Easy victories and uphill battles in coreference resolution, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1971–1982.
[43] G. Durrett, D. Klein, A joint model for entity analysis: coreference, typing, and linking, Trans. Associat. Comput. Linguist. 2 (2014) 477–490.
[44] E.R. Fernandes, C.N. Dos Santos, R.L. Milidiú, Latent structure perceptron with feature induction for unrestricted coreference resolution, in: Joint Conference on EMNLP and CoNLL-Shared Task, Association for Computational Linguistics, 2012, pp. 41–48.
[45] C.J. Fillmore, Pragmatically controlled zero anaphora, in: Annual Meeting of the Berkeley Linguistics Society, 12, 1986, pp. 95–107.
[46] J.R. Finkel, C.D. Manning, Enforcing transitivity in coreference resolution, in: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Association for Computational Linguistics, 2008, pp. 45–48.
[47] C. Gasperin, T. Briscoe, Statistical anaphora resolution in biomedical texts, in: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics, 2008, pp. 257–264.
[48] N. Ge, J. Hale, E. Charniak, A statistical approach to anaphora resolution, Sixth Workshop on Very Large Corpora, 1998.
[49] A. Ghaddar, P. Langlais, Wikicoref: an english coreference-annotated corpus of wikipedia articles., LREC, 2016.
[50] A.E. Goldberg, L.A. Michaelis, One among many: anaphoric one and its relationship with numeral one, Cogn. Sci. 41 (S2) (2017) 233–258.

[51] R. Grishman, B. Sundheim, Message understanding conference-6: a brief history, in: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, 1, 1996.
[52] D. Gross, J. Allen, D. Traum, The trains 91 dialogues (trains tech. note 92–1), Rochester, NY: University of Rochester, Department of Computer Science, 1993.
[53] B.J. Grosz, S. Weinstein, A.K. Joshi, Centering: a framework for modeling the local coherence of discourse, Comput. Linguist. 21 (2) (1995) 203–225.
[54] L. Guillou, C. Hardmeier, A. Smith, J. Tiedemann, B. Webber, Parcor 1.0: a parallel pronoun-coreference corpus to support statistical mt, in: 9th International Conference on Language Resources and Evaluation (LREC), MAY 26–31, 2014, Reykjavik, ICELAND, European Language Resources Association, 2014, pp. 3191–3198.
[55] J. Gundel, N. Hedberg, R. Zacharski, Pronouns without np antecedents: how do we know when a pronoun is referential, Anaphora Process (2005) 351–364.
[56] A. Haghighi, D. Klein, Unsupervised coreference resolution in a nonparametric bayesian model, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 848–855.
[57] A. Haghighi, D. Klein, Simple coreference resolution with rich syntactic and semantic features, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3, Association for Computational Linguistics, 2009, pp. 1152–1161.
[58] A. Haghighi, D. Klein, Coreference resolution in a modular, entity-centered model, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 385–393.
[59] S.M. Harabagiu, R.C. Bunescu, S.J. Maiorano, Text and knowledge mining for coreference resolution, in: Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, Association for Computational Linguistics, 2001, pp. 1–8.
[60] S.M. Harabagiu, S.J. Maiorano, Knowledge-lean coreference resolution and its relation to textual cohesion and coherence, Relat. Discour./Dialogue Struct. Ref. (1999).
[61] C. Hardmeier, M. Federico, Modelling pronominal anaphora in statistical machine translation, in: IWSLT (International Workshop on Spoken Language Translation); Paris, France; December 2nd and 3rd, 2010., 2010, pp. 283–289.
[62] L. Hasler, C. Orasan, Do coreferential arguments make event mentions coreferential, in: Proc. the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), Citeseer, 2009.
[63] P.A. Heeman, J.F. Allen, The TRAINS 93 Dialogues., Technical Report, ROCHESTER UNIV NY DEPT OF COMPUTER SCIENCE, 1995.
[64] I. Heim, The semantics of definite and indefinite NPs, University of Massachusetts at Amherst Dissertation (1982).
[65] J.R. Hobbs, Resolving pronoun references, Lingua 44 (4) (1978) 311–338.
[66] Y. Hou, K. Markert, M. Strube, Global inference for bridging anaphora resolution, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 907–917.
[67] N. Jakob, I. Gurevych, Using anaphora resolution to improve opinion target identification in movie reviews, in: Proceedings of the ACL 2010 Conference Short Papers, Association for Computational Linguistics, 2010, pp. 263–268.
[68] C. Kennedy, B. Boguraev, Anaphora for everyone: pronominal anaphora resoluation without a parser, in: Proceedings of the 16th Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics, 1996, pp. 113–118.
[69] R. Kibble, A reformulation of rule 2 of centering theory, Comput. Linguist. 27 (4) (2001) 579–587.
[70] J.-D. Kim, T. Ohta, Y. Tateisi, J. Tsujii, Genia corpus—a semantically annotated corpus for bio-textmining, Bioinformatics 19 (suppl_1) (2003) i180–i182.
[71] J.-D. Kim, T. Ohta, J. Tsujii, Corpus annotation for mining biomedical events from literature, BMC Bioinformat. 9 (1) (2008) 10.
[72] J.-D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, J. Tsujii, Overview of bionlp shared task 2011, in: Proceedings of the BioNLP shared task 2011 workshop, Association for Computational Linguistics, 2011, pp. 1–6.
[73] S. Lappin, H.J. Leass, An algorithm for pronominal anaphora resolution, Comput. Linguist. 20 (4) (1994) 535–561.
[74] T.T. Le, T.H. Vo, D.T. Mai, T.T. Quan, T.T. Phan, Sentiment analysis using anaphoric coreference resolution and ontology inference, in: International Workshop on Multi-Disciplinary Trends in Artificial Intelligence, Springer, 2016, pp. 297–303.
[75] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, D. Jurafsky, Deterministic coreference resolution based on entity-centric, precision-ranked rules, Comput. Linguist. 39 (4) (2013) 885–916.
[76] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, D. Jurafsky, Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task, in: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, Association for Computational Linguistics, 2011, pp. 28–34.
[77] H. Lee, M. Recasens, A. Chang, M. Surdeanu, D. Jurafsky, Joint entity and event coreference resolution across documents, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, 2012, pp. 489–500.
[78] H. Lee, M. Surdeanu, D. Jurafsky, A scaffolding approach to coreference resolution integrating statistical and rule-based models, Nat. Lang. Eng. 23 (5) (2017) 733–762.
[79] K. Lee, L. He, M. Lewis, L. Zettlemoyer, End-to-end neural coreference resolution., 2017 arXiv:1707.07045.
[80] T. Liang, D.-S. Wu, Automatic pronominal anaphora resolution in english texts, Int. J. Comput. Linguist. Chinese Lang. Process. Volume 9, Number 1, February 2004: Special Issue on Selected Papers from ROCLING XV 9 (1) (2004) 21–40.
[81] J. Lu, D. Venugopal, V. Gogate, V. Ng, Joint inference for event coreference resolution, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 3264–3275.
[82] X. Luo, On coreference resolution performance metrics, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2005, pp. 25–32.
[83] C. Ma, J.R. Doppa, J.W. Orr, P. Mannem, X. Fern, T. Dietterich, P. Tadepalli, Prune-and-score: Learning for greedy coreference resolution, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 2115–2126.
[84] Y. Ma, H. Peng, E. Cambria, Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM, in: AAAI, 2018, pp. 5876–5883.
[85] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, A. Gelbukh, Sentiment and sarcasm classification with multitask learning, IEEE Intell. Syst. 34 (3) (2019) 38–43.
[86] A. Marasović, L. Born, J. Opitz, A. Frank, A mention-ranking model for abstract anaphora resolution., 2017 arXiv:1706.02256.
[87] A. McCallum, B. Wellner, Object consolidation by graph partitioning with a conditionally-trained distance metric, KDD Workshop on Data Cleaning, Record Linkage and Object Consolidation, Citeseer, 2003.
[88] A. McCallum, B. Wellner, Conditional models of identity uncertainty with application to noun coreference, in: Advances in Neural Information Processing Systems, 2005, pp. 905–912.
[89] J.F. McCarthy, W.G. Lehnert, Using decision trees for coreference resolution., 1995 cmp-lg/9505043.
[90] Y. Mehta, N. Majumder, A. Gelbukh, E. Cambria, Recent trends in deep learning based personality detection, Artif. Intell. Rev. (2020).
[91] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[92] R. Mitkov, Robust pronoun resolution with limited knowledge, in: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2, Association for Computational Linguistics, 1998, pp. 869–875.
[93] R. Mitkov, Anaphora resolution: the state of the art, Citeseer, 1999.
[94] R. Mitkov, Outstanding issues in anaphora resolution, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2001, pp. 110–125.
[95] R. Mitkov, Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems, Appl. Artif. Intell. 15 (3) (2001) 253–276.
[96] R. Mitkov, Anaphora resolution, Routledge, 2014.
[97] R. Mitkov, R. Evans, C. Orasan, A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2002, pp. 168–186.
[98] N.S. Moosavi, M. Strube, Which coreference evaluation metric do you trust? Aproposal for a link-based entity aware metric, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1, 2016, pp. 632–642.
[99] N.S. Moosavi, M. Strube, Lexical features in coreference resolution: To be used with caution, arXiv preprint arXiv:1704.06779 (2017).
[100] V. Ng, Machine learning for coreference resolution: from local classification to global ranking, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2005, pp. 157–164.
[101] V. Ng, Supervised noun phrase coreference research: The first fifteen years, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 1396–1411.
[102] V. Ng, Machine learning for entity coreference resolution: aretrospective look at two decades of research., in: AAAI, 2017, pp. 4877–4884.
[103] V. Ng, C. Cardie, Combining sample selection and error-driven pruning for machine learning of coreference rules, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10, Association for Computational Linguistics, 2002, pp. 55–62.
[104] V. Ng, C. Cardie, Improving machine learning approaches to coreference resolution, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 104–111.
[105] C. Nicolae, G. Nicolae, Bestcut: a graph algorithm for coreference resolution, in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2006, pp. 275–283.
[106] N. Nicolov, F. Salvetti, S. Ivanova, Sentiment analysis: does coreference matter, in: AISB 2008 Convention Communication, Interaction and Social Intelligence, 1, 2008, p. 37.
[107] B. O'Connor, M. Heilman, Arkref: A rule-based coreference resolution system, arXiv preprint arXiv:1310.1975 (2013).
[108] L. Oneto, F. Bisio, E. Cambria, D. Anguita, Statistical learning theory and ELM for big social data analysis, IEEE Comput. Intell. Mag. 11 (3) (2016) 45–55.
[109] H. Peng, D. Khashabi, D. Roth, Solving hard coreference problems, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 809–819.
[110] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[111] M. Poesio, Discourse annotation and semantic annotation in the gnome corpus, in: Proceedings of the 2004 ACL Workshop on Discourse Annotation, Association for Computational Linguistics, 2004, pp. 72–79.

[112] M. Poesio, R. Artstein, et al., Anaphoric annotation in the arrau corpus., LREC, 2008.
[113] M. Poesio, M.A. Kabadjov, A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation., LREC, 2004.
[114] H. Poon, P. Domingos, Joint unsupervised coreference resolution with markov logic, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 650–659.
[115] S. Pradhan, X. Luo, M. Recasens, E. Hovy, V. Ng, M. Strube, Scoring coreference partitions of predicted mentions: a reference implementation, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, 2014, NIH Public Access, 2014, p. 30.
[116] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, Y. Zhang, Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes, in: Joint Conference on EMNLP and CoNLL-Shared Task, Association for Computational Linguistics, 2012, pp. 1–40.
[117] S. Pradhan, L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, N. Xue, Conll-2011 shared task: Modeling unrestricted coreference in ontonotes, in: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, Association for Computational Linguistics, 2011, pp. 1–27.
[118] S. Preuss, Anaphora resolution in machine translation, TU, Fachbereich 20, Projektgruppe KIT, 1992.
[119] J. Pustejovsky, J. Castano, R. Sauri, A. Rumshinsky, J. Zhang, W. Luo, Medstract: creating large-scale information servers for biomedical libraries, in: Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain-Volume 3, Association for Computational Linguistics, 2002, pp. 85–92.
[120] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[121] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, C. Manning, A multi-pass sieve for coreference resolution, in: Proceedings of 2010 Conference on Empirical Methods in NAturalLanguage Processing, Association for Computational Linguistics, 2010, pp. 492–501.
[122] A. Rahman, V. Ng, Supervised models for coreference resolution, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, Association for Computational Linguistics, 2009, pp. 968–977.
[123] A. Rahman, V. Ng, Coreference resolution with world knowledge, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics, 2011, pp. 814–824.
[124] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. 66 (336) (1971) 846–850.
[125] M. Recasens, E. Hovy, Blanc: implementing the rand index for coreference evaluation, Nat. Lang. Eng. 17 (4) (2011) 485–510.
[126] M. Recasens, M.-C. de Marneffe, C. Potts, The life and death of discourse entities: identifying singleton mentions, in: Proceedings of the 2013 Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies, 2013, pp. 627–633.
[127] M. Recasens, L. Màrquez, E. Sapena, M.A. Martí, M. Taulé, V. Hoste, M. Poesio, Y. Versley, Semeval-2010 task 1: Coreference resolution in multiple languages, in: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2010, pp. 1–8.
[128] M. Recasens, M.A. Martí, Ancora-co: coreferentially annotated corpora for spanish and catalan, Lang. Resour. Eval. 44 (4) (2010) 315–345.
[129] M. Recasens, M. Vila, On paraphrase and coreference, Comput. Linguist. 36 (4) (2010) 639–647.
[130] M. Regneri, R. Wang, Using discourse information for paraphrase extraction, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, 2012, pp. 916–927.
[131] C. Roberts, Modal subordination and pronominal anaphora in discourse, Linguist. Philos. 12 (6) (1989) 683–721.
[132] R.A. Van der Sandt, Presupposition projection as anaphora resolution, J. Semant. 9 (4) (1992) 333–377.
[133] I. Segura-Bedmar, M. Crespo, C. de Pablo, P. Martínez, Drugnerar: linguistic rule-based anaphora resolver for drug-drug interaction extraction in pharmacological documents, in: Proceedings of the Third International Workshop on Data and Text Mining in Bioinformatics, ACM, 2009, pp. 19–26.
[134] W.M. Soon, H.T. Ng, D.C.Y. Lim, A machine learning approach to coreference resolution of noun phrases, Comput. Linguist. 27 (4) (2001) 521–544.
[135] R. Soricut, D. Marcu, Sentence level discourse parsing using syntactic and lexical information, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, Association for Computational Linguistics, 2003, pp. 149–156.
[136] J. Steinberger, M. Poesio, M.A. Kabadjov, K. Ježek, Two uses of anaphora resolution in summarization, Inf. Process. Manag. 43 (6) (2007) 1663–1680.
[137] V. Stoyanov, C. Cardie, N. Gilbert, E. Riloff, D. Buttler, D. Hysom, Coreference resolution with reconcile, in: Proceedings of the ACL 2010 Conference Short Papers, Association for Computational Linguistics, 2010, pp. 156–161.
[138] M. Strube, S. Rapp, C. Müller, The influence of minimum edit distance on reference resolution, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10, Association for Computational Linguistics, 2002, pp. 312–319.
[139] J. Su, X. Yang, H. Hong, Y. Tateisi, J. Tsujii, Coreference resolution in biomedical texts: a machine learning approach, in: Dagstuhl Seminar Proceedings, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2008.
[140] Y. Tateisi, A. Yakushiji, T. Ohta, J. Tsujii, Syntax annotation for the genia corpus, in: Companion Volume to the Proceedings of Conference Including Posters/Demos and Tutorial Abstracts, 2005.
[141] J.R. Tetreault, Analysis of syntax-based pronoun resolution methods, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, 1999, pp. 602–605.
[142] J.R. Tetreault, A corpus-based evaluation of centering and pronoun resolution, Comput. Linguist. 27 (4) (2001) 507–520.
[143] J. Tourille, Extracting Clinical Event Timelines: Temporal Information Extraction and Coreference Resolution in Electronic Health Records, Université Paris-Saclay, 2018 Ph.D. thesis.
[144] O. Uryupina, M. Poesio, C. Giuliano, K. Tymoshenko, Disambiguation and Filtering Methods in Using Web Knowledge for Coreference Resolution, in: Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches, IGI Global, 2012, pp. 185–201.
[145] H. Vala, A. Piper, D. Ruths, The more antecedents, the merrier: resolving multi-antecedent anaphors, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1, 2016, pp. 2287–2296.
[146] A. Valdivia, V. Luzón, E. Cambria, F. Herrera, Consensus vote models for detecting and filtering neutrality in sentiment analysis, Inf. Fus. 44 (2018) 126–135.
[147] Y. Versley, S.P. Pozetto, M. Poesio, V. Eidlman, A. Jern, J. Smith, X. Yang, A. Moschitti, Bart: a modular toolkit for coreference resolution, in: Proceedings of the 46th Annual MEeting of Association for Computational Linguistics on Human Language Technologies: Demo Session, Association for Computational Linguistics, 2008, pp. 9–12.
[148] J.L. Vicedo, A. Ferrández, Importance of pronominal anaphora resolution in question answering systems, in: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2000, pp. 555–562.
[149] R. Vieira, S. Salmon-Alt, C. Gasperin, E. Schang, G. Othero, Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus, Anaphora Process. (2005) 385–403.
[150] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, L. Hirschman, A model-theoretic coreference scoring scheme, in: Proceedings of the 6th Conference on Message Understanding, Association for Computational Linguistics, 1995, pp. 45–52.
[151] M.A. Walker, Evaluating discourse processing algorithms, in: Proceedings of the 27th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 1989, pp. 251–261.
[152] M.A. Walker, A.K. Joshi, E.F. Prince, Centering Theory in Discourse, Oxford University Press, 1998.
[153] K.A. Watson-Gegeo, The pear stories: cognitive, cultural, and linguistic aspects of narrative production, 1981.
[154] L.M. Werlen, A. Popescu-Belis, Using coreference links to improve spanish–to-english machine translation, in: Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017), 2017, pp. 30–40.
[155] J. Weston, A. Bordes, S. Chopra, A.M. Rush, B. van Merriënboer, A. Joulin, T. Mikolov, Towards aicomplete question answering: A set of prerequisite toy tasks, arXiv preprint arXiv:1502.05698 (2015).
[156] S. Wiseman, A.M. Rush, S. Shieber, J. Weston, Learning anaphoricity and antecedent ranking features for coreference resolution, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 1, 2015, pp. 1416–1426.
[157] S. Wiseman, A.M. Rush, S.M. Shieber, Learning global features for coreference resolution., 2016 arXiv:1604.03035.
[158] R. Witte, S. Bergler, Fuzzy coreference resolution for summarization, in: Proceedings of 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization (ARQAS), 2003, pp. 43–50.
[159] X. Yang, J. Su, C.L. Tan, A twin-candidate model for learning-based anaphora resolution, Computat. Linguist. 34 (3) (2008) 327–356.
[160] X. Yang, G. Zhou, J. Su, C.L. Tan, Coreference resolution using competition learning approach, in: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, Association for Computational Linguistics, 2003, pp. 176–183.
[161] X. Yang, G. Zhou, J. Su, C.L. Tan, Improving noun phrase coreference resolution by matching strings, in: International Conference on Natural Language Processing, Springer, 2004, pp. 22–31.
[162] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, M. Huang, Augmenting end-to-end dialogue systems with commonsense knowledge, in: AAAI, 2018, pp. 4970–4977.
[163] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing, IEEE Comput. Intell. Mag. 13 (3) (2018) 55–75.
[164] A. Zeldes, The gum corpus: creating multilayer resources in the classroom, Lang. Resour. Eval. 51 (3) (2017) 581–612.
[165] A. Zeldes, S. Zhang, When annotation schemes change rules help: aconfigurable approach to coreference resolution beyond ontonotes, in: Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016), 2016, pp. 92–101.
[166] L. Zhuang, F. Jing, X.-Y. Zhu, Movie review mining and summarization, in: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, ACM, 2006, pp. 43–50.