ANNUAL REVIEWS

Annual Review of Linguistics

Computational Models of Anaphora

Massimo Poesio,[1] Juntao Yu,[2] Silviu Paun,[1] Abdulrahman Aloraini,[1] Pengcheng Lu,[1] Janosch Haber,[1] and Derya Cokal[3]

[1]Computational Linguistics Lab, Queen Mary University of London, London, United Kingdom; email: m.poesio@qmul.ac.uk

[2]Natural Language and Information Processing Lab, University of Essex, Colchester, United Kingdom

[3]Prominence in Language SFB, University of Cologne, Cologne, Germany

Annu. Rev. Linguist. 2023. 9:28.1–28.27

The *Annual Review of Linguistics* is online at linguistics.annualreviews.org

https://doi.org/10.1146/annurev-linguistics-031120-111653

Keywords

anaphora, anaphora resolution, coreference, corpora for anaphora, neural models for coreference, bridging reference, discourse deixis, Winograd Schema Challenge

Abstract

Interpreting anaphoric references is a fundamental aspect of our language competence that has long attracted the attention of computational linguists. The appearance of ever-larger anaphorically annotated data sets covering more and more anaphoric phenomena in ever-greater detail has spurred the development of increasingly more sophisticated computational models; as a result, the most recent state-of-the-art neural models are able to achieve impressive performance by leveraging linguistic, lexical, discourse, and encyclopedic information. This article provides a thorough survey of anaphora resolution (coreference) throughout this development, reviewing the available data sets and covering both the preneural history of the field and—in more detail—current neural models, including research on less-studied aspects of anaphoric interpretation such as bridging reference resolution and discourse deixis interpretation.

1. INTRODUCTION

Interpreting anaphoric references is an aspect of our linguistic competence that has attracted much interest from theoretical, psycho-, and computational linguists, in part because it straddles sentential and intersentential interpretation; in part because it draws on all types of information, from lexical to syntactic to contextual information to commonsense knowledge; and in part, finally, because human judgments on anaphoric interpretation are much sharper than judgments on aspects of interpretation such as rhetorical structure or even syntax. Evidence from anaphoric reference has played a key role in the development of modern theories of syntax (e.g., binding; Büring 2005), of discourse models and their role in semantics (Karttunen 1976, Webber 1979, Heim 1982, Kamp et al. 2011), and of salience and its role in interpretation [Sidner 1979, Grosz & Sidner 1986, Grosz et al. 1995 (1986)].

Anaphoric reference is also one of the most active areas of computational linguistics (CL). The study of computational models of anaphora underwent several paradigm shifts, transitioning from cognitively and linguistically inspired models (Hobbs 1978, Sidner 1979, Carter 1987, Hobbs et al. 1993, Lappin & Leass 1994, Poesio 1994) to data-driven and statistical models (Aone & Bennett 1995, Humphreys et al. 1998, Vieira & Poesio 2000, Soon et al. 2001, Ng & Cardie 2002b, Daume & Marcu 2005, Ponzetto & Strube 2006, Denis & Baldridge 2007, Bengtson & Roth 2008, Rahman & Ng 2011) when the first annotated data sets appeared (Chinchor & Sundheim 1995, Doddington et al. 2000). Once more substantial annotated data sets started to become available (Hinrichs et al. 2005, Pradhan et al. 2012, Poesio & Artstein 2008, Nedoluzhko et al. 2009, Recasens & Martí 2010, Muzerelle et al. 2014, Poesio et al. 2019, Uryupina et al. 2020, Zeldes 2020), research in this area boomed, leading first to advanced statistical models (Björkelund & Kuhn 2014, Durrett & Klein 2014, Fernandes et al. 2014, Martschat & Strube 2015) and then to neural models that appear to have addressed many of the problems of previous models, resulting in an impressive performance (Wiseman et al. 2015; Lee et al. 2017, 2018; Joshi et al. 2020; Yu et al. 2020c). In the last 20 years, there has also been progress in the creation of data sets for genres other than news (Yang et al. 2004, Cohen et al. 2017, Bamman et al. 2020) as well as benchmarks for testing aspects of anaphoric interpretation not properly tested with full-document data sets (Levesque et al. 2012, Webster et al. 2018).

The first author of this review coauthored a book on anaphora resolution fairly recently (Poesio et al. 2016b). However, CL moves fast. By the time that book was completed, the field had already changed dramatically with the appearance of the first neural models. In addition, that book did not cover aspects of anaphoric reference that had not been extensively studied in CL until very recently, such as bridging reference, discourse deixis, and the interpretation of plurals (Hou et al. 2013, Marasović et al. 2017, Hou et al. 2018, Roesiger et al. 2018, Yu & Poesio 2020, Yu et al. 2021). This review thus aims to provide a more complete (if more succinct) survey of the area, including those developments not covered in the book by Poesio et al. (2016b).

2. THE COMPUTATIONAL PERSPECTIVE ON ANAPHORA

In this review, we do not attempt to provide a full introduction to the linguistics and psycholinguistics of anaphora, which are well covered in works by, for instance, Kamp & Reyle (1993), Garnham (2001), Büring (2005), and Gundel & Abbott (2019) as well as Poesio et al. (2016b). We instead concentrate on the aspects of anaphora most studied in CL.

2.1. The Linguistics of Anaphora

2.1.1. Anaphoric expression and discourse models. Most natural language expressions depend to some extent on context for their interpretation. Anaphoric expressions are characterized

by depending in part or entirely on the entities mentioned in the linguistic context—the previous utterances and their content. Such dependency is particularly obvious in the case of pronouns like *he*, *him*, or *his* in the following text, whose interpretation entirely depends on which entity is chosen as the antecedent. But other types of noun phrases (NPs) depend on the linguistic context for their interpretation as well, including nominals such as *your father* or even proper nouns like the second instance of *Maupin* in example 1, which could refer either to Armistead Jones Maupin Jr. (the author of *Tales of the City*) or his father, Armistead Jones Maupin (who served in the US Navy):

(1) [Maupin]$_i$ recalls [his]$_i$ mother trying to shield [him]$_i$ from [[his]$_i$ father's]$_j$ excesses.
"[[Your]$_i$ father]$_j$ doesn't mean it," she would console [him]$_i$.
"[He]$_j$ loves [you]$_i$, [he]$_j$'s a good man."
And for years [he]$_i$ thought she was making excuses.
"But she wasn't. [He]$_j$ is a good man." Just a product of [his]$_j$ time.

Most computational models of anaphoric reference interpretation (a task we refer to here as anaphora resolution) tend to be based on (some version of) the discourse model approach to the semantics of anaphora pioneered by Bransford and colleagues in psycholinguistics (for a review, see Garnham 2001) and by Karttunen (1976) in theoretical linguistics and Webber (1979) in CL, which led to dynamic semantics theories (Kamp et al. 2011). In such models, interpretation takes place against a context that consists of discourse entities; each new sentence may contain references to these discourse entities and/or result in new entities being added to the context. In CL, discourse entities typically take the form of coreference chains: clusters of mentions all referring to the same entity.

2.1.2. Anaphora and coreference. Early work focused primarily on pronominal anaphoric reference, but ever since the appearance of the first substantial anaphorically annotated corpora and in particular since the first classic model of coreference resolution (Soon et al. 2001), most research has been concerned with developing models capable of interpreting all types of reference to discourse entities via nominals. Yet, in much CL/natural language processing literature a distinction is still made between anaphora resolution and coreference resolution, and the term anaphora is used to indicate pronominal anaphora only. In this review, the terms anaphora and anaphoric reference are used in the more general sense of reference to entities in the discourse model used in semantics (see, e.g., Lyons 1977, Kamp & Reyle 1993) and psycholinguistics (see, e.g., Garnham 2001). In Discourse Representation Theory (DRT) (Kamp et al. 2011), for instance, the proper name *Maupin* in example 1 adds to the discourse model a new discourse entity *i*, and all subsequent mentions of Maupin, whether using pronouns or proper names, are interpreted as anaphoric references to entity *i*.

2.1.3. The semantic function of noun phrases. Referring NPs introduce new entities in a discourse or link to previously introduced entities; examples include the references to Maupin in example 1. The items annotated in anaphoric corpora tend to be a subset of referring NPs. But other types of NPs also exist. Quantificational NPs such as *No one* in *No one would put the blame on him/herself* (Partee 1972) do not refer to an individual or set of individuals but can still participate in anaphoric relations even though anaphoric reference to quantifiers has distinctive properties (Partee 1972) and is subject to semantic constraints (Karttunen 1976). Predicative NPs express properties of objects: For instance, in the clause *He is a good man* in example 1, the NP *a good man* does not introduce a new discourse entity or refer back to an existing discourse entity but instead expresses a property of Maupin's father. Finally, in languages like English, forms like *it* and *there* can also be used to express semantically vacuous expletives as well as pronouns, as in *It is half*

past two. There is substantial disagreement in CL on whether all types of NPs or only referring expressions should be annotated in a corpus for anaphora (Poesio et al. 2016a).

2.1.4. Incorporated and zero anaphora. In languages other than English, anaphoric reference can be expressed implicitly, or the anaphora can be incorporated in a nonnominal constituent such as a verb. A great deal of attention has been paid in CL to the identification and interpretation of zero anaphora—anaphoric references in which a verbal argument is not realized, which occur for languages such as Arabic, Chinese, Italian, Japanese, and Spanish.

2.1.5. Constraints on anaphoric interpretation. Syntactic (Büring 2005) and semantic (Karttunen 1976, Heim 1982, Kamp & Reyle 1993) constraints on anaphora have played an important role in linguistic theorizing but only a limited one in recent computational models of anaphora. On the other end, there has been extensive work on the pragmatic effects of discourse structure on anaphoric reference, which is briefly discussed in Section 2.2.

2.1.6. Associative anaphora (bridging). Most computational models of anaphora focus on identity relations, largely because of the coverage of existing data sets (see Section 3). However, there has been much interest in associative anaphora as well (Clark 1977), where the anaphoric expression is related to its antecedent by a relation other than identity, as in example 2, in which the kitchen and the garden are associated with the flat introduced in the first sentence. This type of anaphoric reference is usually called a bridging reference in CL because a bridging inference is generally required to identify the antecedent (Clark 1977):

(2) We saw [a flat]$_i$ yesterday. [The kitchen]^i_j is spacious but [the garden]^i_k is very small.

2.1.7. Other cases of anaphoric reference to antecedents not explicitly introduced with nominals. Other cases of anaphoric reference to antecedents not introduced via nominals have also been studied in CL (Eschenbach et al. 1989, Webber 1991, Kolhatkar et al. 2018). One is discourse deixis, or anaphora with nonnominal antecedents (Webber 1991, Kolhatkar et al. 2018; see example 3). This is a type of anaphora in which the antecedent is an abstract entity associated with the propositional content of a segment:

(3) The municipal council had to decide [whether to balance the budget by raising revenue or cutting spending]$_i$. The council had to come to a resolution by the end of the month. [This issue]$_i$ was dividing communities across the country.

Interpreting some cases of anaphoric reference requires updating the context via some explicit interpretation. The simplest among these are the cases of split antecedent anaphora studied by Eschenbach et al. (1989) and Kamp & Reyle (1993) and illustrated in example 4. The antecedent for *they* is a plural entity that is not explicitly mentioned but somehow constructed out of the explicitly mentioned Michael and Maria:

(4) [Michael]$_i$ was at the cinema with [Maria]$_j$. [They]$_{i+j}$ had a great time.

2.2. Factors Affecting Anaphoric Interpretation

2.2.1. Agreement constraints. Agreement constraints such as gender and number are one of the strongest types of disambiguation factors (Garnham 2001). For instance, the pronoun *she* in the second sentence in example 1 is unambiguous because there is only one entity of feminine gender in the example. However, such constraints cannot always be relied upon. Real text contains cases

like example 5, where *its* is erroneously used to refer to *a customer*. And the gender of entities is not always known. Agreement mismatch problems are also becoming more common because of the increasing use of plural pronouns to avoid gender bias, as in example 6:

(5) to get [a customer's]$_i$ 1100 parcel-a-week load to [its]$_i$ doorstep

(6) when [the doctor]$_i$ arrives, ask [them]$_i$ about your cough

2.2.2. Lexical and commonsense knowledge. Lexical and commonsense knowledge can be an equally strong disambiguation factor. One of the best-known illustrations of this is the minimal pair (example 7) (Winograd 1972). The only difference between examples 7a and 7b is the verb in the second clause, but that change shifts the preferred interpretation for *they* from *the council* in example 7a to *the women* in example 7b:

(7a) [The city council]$_i$ refused [the women]$_j$ [a permit]$_k$ because [they]$_i$ feared violence.
(7b) [The city council]$_i$ refused [the women]$_j$ [a permit]$_k$ because [they]$_j$ advocated violence.

This minimal pair recently acquired great prominence as the first example of what has become known as the Winograd Schema approach to evaluating anaphora resolution proposed by Levesque et al. (2012) (see Section 3).

2.2.3. Syntactic constraints. The prohibition for *him* to corefer with *John* in **John$_i$ likes him$_i$* played an important role in linguistic theorizing, as discussed above (Büring 2005). Such constraints also played an important role in early models of pronominal interpretation such as Hobbs's "naive algorithm" (Hobbs 1978) but not in recent models.

2.2.4. Discourse factors. It has long been known that more recently introduced entities are more likely antecedents; in CL, Hobbs (1978), for instance, reported that in his corpus, 98% of pronoun antecedents were in the current or the previous sentence. A stronger hypothesis is that linguistic focusing mechanisms—attentional mechanisms of the type found in visual interpretation—also affect the interpretation of anaphoric expressions (Grosz 1977, Sidner 1979, Sanford & Garrod 1981). According to the best-known theory of this type in CL, proposed by Grosz & Sidner (1986), two levels of structure exist in discourse: the global focus, which specifies the articulation of discourse segments; and the local focus, which specifies how, utterance by utterance, the relative salience of entities changes. Authors such as Grosz & Sidner (1986), Mann & Thompson (1988), Webber (1991), and Asher & Lascarides (2003) have argued that discourse segments have a hierarchical structure that affects anaphoric interpretation (see, e.g., Fox 1987 for an analysis of some of these claims). Sidner (1979) proposed the first detailed theory of the local focus; Centering [Grosz et al. 1995 (1986)] eventually evolved into the dominant theory of the local focus in CL and, to some extent, in psycholinguistics (Walker et al. 1998, Poesio et al. 2004b).

2.3. Ambiguity

One property of anaphoric reference that was not extensively studied in either the linguistic or the psycholinguistic literature on anaphora but that has been highlighted from large-scale anaphoric annotation efforts in CL is the fact that many anaphoric expressions do not have a preferred interpretation in context (Poesio & Artstein 2005, Recasens et al. 2011). The prevalence of ambiguous cases in anaphorically annotated corpora ranges from 10–15% in more formal texts (Pradhan et al. 2012, Poesio et al. 2019) to 30–40% in dialogue data and when discourse deixis is also annotated (Poesio & Artstein 2005). This evidence suggests that for a proper empirical

investigation of anaphoric reference, multiple interpretations should be preserved, but this is seldom done in existing data sets (see, however, Section 3).

A more extensive discussion of the CL perspective on the linguistic and cognitive aspects of anaphora can be found in previous work by Poesio (2016) and in earlier monographs (e.g., Mitkov 2002).

3. THE DATA: FULL-TEXT CORPORA AND BENCHMARKS

CL is very much driven by the availability of data sets. Early anaphora resolution was developed and evaluated on the basis of individual examples or at best on the basis of collections of "hard" examples (also known as benchmarks) (see, e.g., Carter 1987). This all changed when the first full-text corpora became available (Chinchor & Sundheim 1995, Doddington et al. 2000) and turned anaphora resolution into a data-driven field. In this section we discuss the main developments regarding data set availability in the field of anaphora resolution.

3.1. Annotating Nominal Anaphora: The Options

3.1.1. The definition of markable. The prototypical markable—the item to annotate—in most anaphoric data sets is the NP, generally considered in its entirety; however, some differences exist between data sets, and other sentence constituents are also considered, such as possessive pronouns as well as zeros for languages such as Arabic, Chinese, Italian, and Japanese.

Semantic and discourse restrictions on the definition of markable are often also imposed. In particular, very few corpora attempt to annotate all types of NPs discussed in Section 2.1 (Poesio et al. 2016a). So, for instance, in the most used anaphoric data set for Arabic, Chinese, and English—OntoNotes (Pradhan et al. 2012)—only some types of referring NPs and some types of predicative NPs are annotated (see Section 2.1); other types of predicative NPs, expletives, and other types of nonreferring NPs are not. In fact, in OntoNotes and other data sets, only NPs that refer to entities mentioned more than once are annotated; so-called singletons are not. As a result, most CL work on anaphora resolution focuses on referring expressions only and does not attempt to resolve ambiguities such as those between the expletive or anaphoric interpretation of *it* and the predicative or referential interpretation of some indefinite NP. As an additional restriction, some data sets created to study the effect of anaphora on information extraction only annotate markables denoting certain types of entities; for instance, in the ACE corpora (Doddington et al. 2000), only NPs that refer to a few types of entities are annotated (e.g., persons, organizations), and others are not (e.g., references to animals, art objects, substances).

3.1.2. Predication. One of the most discussed properties of the annotation schemes used for the original information-extraction-led data sets such as MUC and ACE (Chinchor & Sundheim 1995, Doddington et al. 2000) was the inclusion in "coreference resolution" of what linguistically would be considered cases of predication. In these corpora, *a good man* would be marked as coreferring with Maupin's father in the third sentence of example 1. This approach raised the problems discussed by, among others, van Deemter & Kibble (2000), leading, for instance, to implausible coreference relations when predications change over time, such as net income in example 8 (from the WSJ portion of the ARRAU corpus). Contemporary corpora greatly differ with respect to how they treat predication (for more discussion, see Zeldes 2022):

(8) [The company] said net was [38 cents] a share in its fiscal-first quarter ended Sept. 30, from [35 cents] a share a year ago.

Table 1 The most widely used anaphoric data sets

Name	Language	Genre(s)	Size	TB?	AnnoS	NR?	BR?	DD?	A?
ACE-2005	English	News	500K	✓	MUC				
ANCOR	French	Spoken	500K	✓	MATE		✓		
ANCORA	Catalan	News	500K	✓	MATE	✓	✓	✓	
ANCORA	Spanish	News	500K	✓	MATE	✓	✓	✓	
ARRAU	English	Multiple	350K		MATE	✓	✓	✓	✓
COREA	Dutch	News	325K	✓	MATE	✓	✓	✓	
CRAFT	English	Biomed	800K	✓	M/O				
DEMOCRAT	French	Diachronic	300K		MATE				
GENIA	English	Biomed	400K	✓	MUC				
GUM	English	Multiple	130K	✓	MUC	✓	✓	✓	
ISNOTES	English	News	50 docs	✓	M/O		✓		
LITBANK	English	Fiction	200K	✓	M/O				
NAIST	Japan	News	1M	✓	MATE				
ONTONOTES	Arabic	News	300K	✓	M/O			✓	
ONTONOTES	Chinese	News	1.2M	✓	M/O			✓	
ONTONOTES	English	News	1.5M	✓	M/O				
PCC	Polish	News	530K	✓	MATE		✓		
PCEDT	Czech	News	1.15M	✓	MATE	✓	✓	✓	
PCEDT	English	News	1.17M	✓	MATE	✓	✓	✓	
PDT	Czech	News	800K	✓	MATE	✓	✓	✓	
PH.DET.3	English	Multiple	1.2M		MATE	✓			✓
PRECO	English	Learner	10M		MUC				
TÜBA-DZ	German	News	1.56M	✓	MATE				

Abbreviations: A, ambiguity; AnnoS, annotation scheme; BR, bridging references; DD, discourse deixis; NR, nonreferring expressions; TB, Treebanking.

3.1.3. The range of relations. All anaphoric data sets annotate identity—mentioning again a previously mentioned entity—although as we have seen, some data sets only consider identity relations between a subset of the mentions. For many years only small, dedicated data sets were available to study bridging reference resolution, such as GNOME (Poesio et al. 2004a) and ISNOTES (Markert et al. 2012). However, bridging references are annotated in many if not most of the more recent larger data sets (see **Table 1**). But it should be noted that there is much less agreement on the annotation schemes for bridging reference than on those for identity reference (Poesio et al. 2016a, Roesiger et al. 2018). Even smaller is the number of annotation projects that cover discourse deixis, but again the number is growing (see **Table 1**). However, substantial differences exist between the guidelines adopted in these different projects (for details, see Kolhatkar et al. 2018).

3.1.4. Ambiguity. Only a few corpora provide information about cases of anaphoric ambiguity. In ARRAU, ambiguity is marked explicitly—annotators can provide multiple interpretations. In *Phrase Detectives*, ambiguity is marked implicitly—annotators can provide only one interpretation, but because a large number of players provide judgments for each markable (20 on average), disagreements in interpretation can emerge. In ANCORA and the PCC, annotators can use a relation of quasi-identity when coreference is possible but not certain.

3.1.5. Universal Anaphora. To promote standards for annotating and representing multilayer and multilingual anaphorically annotated corpora, the Universal Anaphora initiative was recently launched (**http://www.universalanaphora.org**), modeled on the Universal Dependencies initiative (**http://universaldependencies.org**).

3.2. Full-Text Corpora for Anaphora

3.2.1. Early corpora. The earliest anaphoric data set we are aware of is the IBM/UCREL Anaphoric Treebank (McEnery et al. 1997). This resource was annotated according to a linguistically motivated scheme, arguably the most ambitious anaphoric scheme tried so far, covering not only bridging and discourse deixis but also various types of ellipsis. Unfortunately, however, the resource was never made publicly available. So the data sets that really kick-started the data-driven shift in anaphora resolution were the corpora created for the Message Understanding Conference (MUC) and Automatic Content Extraction (ACE) shared tasks (Chinchor & Sundheim 1995, Doddington et al. 2000). These shared tasks also introduced the coreference task as currently understood, in terms of terminology (e.g., use of the term "mentions" to refer to the items to classify) and of focus on nominal anaphora only. Equally importantly, these shared tasks led to the introduction of the first evaluation metrics designed specifically for anaphora (see Section 5). However, the task definition also raised issues such as the conflation of predication and anaphoric reference discussed earlier, or, in ACE, the restriction on the range of entities considered.

3.2.2. Linguistically motivated data sets. The discussions about the specification of the coreference task in MUC and ACE (van Deemter & Kibble 2000) eventually led to proposals for the annotation of anaphoric information (Passonneau 1997, Poesio et al. 1999) that were more directly based on the linguistic approach to anaphora discussed in Section 2.1. Most of the corpora developed since have adopted a similar approach (Poesio 2004, Hinrichs et al. 2005, Hendrickx et al. 2008, Poesio & Artstein 2008, Nedoluzhko et al. 2009, Recasens & Martí 2010, Pradhan et al. 2012, Ogrodniczuk et al. 2015, Landragin 2016, Zeldes 2020). In particular, the creation of OntoNotes (Pradhan et al. 2012) and the shared tasks based on OntoNotes and other data sets of this type (Recasens et al. 2010, Pradhan et al. 2012) led to a move away from the modeling of coreference in the sense of MUC and ACE and toward anaphora resolution as traditionally conceived in linguistics and psychology.

3.2.3. Genres. Most of the early data sets focused on news articles and broadcasts, but the more recent data sets cover other genres as well. This is important because the news genre provides a skewed picture of the use of anaphoric reference in language; focusing exclusively on such data limits both the generality of the linguistic findings and the usefulness of models trained on the data when applied to other domains (Xia & Durme 2021). Substantial corpora now exist for studying anaphora resolution in the biomedical domain, including, for instance, GENIA (Yang et al. 2004) and CRAFT (Cohen et al. 2017). Two other genres for which substantial data sets have become available include encyclopedic texts, particularly from Wikipedia (covered, e.g., in the *Phrase Detectives* corpus; Poesio et al. 2019), and fiction and literary texts (covered, e.g., in *Phrase Detectives* and LITBANK; Bamman et al. 2020).

3.2.4. Main anaphoric data sets in use today. **Table 1** summarizes the anaphoric data sets most widely used today. Only corpora of at least 300,000 tokens are listed in **Table 1**, with the exception of GUM, ISNOTES (Markert et al. 2012), and LITBANK, which are widely used. For each data set, **Table 1** lists the language; the genre(s); the size in tokens; whether multiple levels of annotation are included (Treebanking); which definition of coreference is used [MUC (Chinchor &

Sundheim 1995), MATE (Poesio et al. 1999), or M/O for the version of the MATE guidelines developed for ONTONOTES (Pradhan et al. 2012)]; whether nonreferring expressions (NR?), bridging references (BR?), and discourse deixis (DD?) are annotated; and whether multiple interpretations for ambiguous markables are included.

3.3. Benchmarks

As discussed above, the first computational models of anaphora were tested against benchmarks containing examples of whichever aspect of anaphoric interpretation a model was developed to handle (Hobbs 1978, Carter 1987). However, this approach was of limited use in assessing the performance of a computational model on real text, so full-text evaluation became the standard approach to evaluation in the field once data sets like MUC became available. But in recent years the realization has been growing that full-text evaluation has limitations too—namely, that because of the prevalence of relatively easy cases in test data sets, a high score may not indicate truly good performance (Barbu & Mitkov 2001, Webster et al. 2018)—the more so when many of the hard cases are excluded a priori because of insufficient agreement (Poesio & Artstein 2005, Recasens et al. 2011). As a result, we are witnessing a return to benchmarks as a way of evaluating anaphora resolution.

3.3.1. The Winograd Schema Challenge. Arguably, the beginning of this trend can be traced back to the proposal by Levesque et al. (2012) of a Winograd Schema Challenge—a benchmark consisting of several hundred minimal pairs based on Winograd's (1972) example (see example 7 above). Since the original paper, many larger data sets have been proposed based on this idea, such as the Definite Pronoun Resolution data set by Rahman & Ng (2012); the Winograd Natural Language Inference (WNLI) data set, a textual entailment version of the Winograd Schema Challenge included in the GLUE data set (Wang et al. 2019); and WINOGRANDE, which consists of 44,000 examples (Sakaguchi et al. 2020).

3.3.2. Resolving gender-ambiguous cases. To test the ability of systems to resolve pronouns without the help of gender cues, the GAP data set was launched (Webster et al. 2018).

3.4. Remaining Gaps

We are in a much better situation than at the beginning of the data-driven era, and quality data sets of a substantial size are now available for many languages. But significant gaps remain. For one thing, many languages are still not covered or are covered only by relatively small data sets; for instance, the largest available data sets for Arabic just pass the 300,000-token threshold used for **Table 1**, and the only data sets we are aware of for some of the most spoken languages in the world—Bengali, Hindi, Portuguese, Russian, and Turkish—do not. Also, the focus so far has been mainly on written language. Very few data sets cover spoken language, and we are aware of only one large corpus of spoken language annotated for coreference: the ANCOR corpus of spoken French (Muzerelle et al. 2014). [A medium-sized dialogue corpus for English was recently created for the CODI/CRAC shared task on anaphora in dialogue (Khosla et al. 2021).] Last but not least, there are still many aspects of anaphoric interpretation (e.g., discourse deixis) for which we lack a solid theoretical foundation. And even our understanding of identity anaphora as reflected in the guidelines used is still partial; Zaenen's admonitions that, for instance, "The problems with the 'coreference' annotation tasks of MUC and the like are well documented and not solved" (Zaenen 2006, p. 578), still apply.

An extensive discussion of anaphoric data sets can be found in a book chapter by Poesio et al. (2016a), but it does not cover data sets released since 2015 (for those, see Nedoluzhko et al. 2021

and the Universal Anaphora pages). The data sets for biomedical information are surveyed by Cohen et al. (2017), and the benchmarks for the Winograd Schema Challenge by Kocijan et al. (2020). Ide & Pustejovsky (2017) provide more in-depth discussion of some corpora.

4. THE PRENEURAL PERIOD

The history of anaphora resolution research can be divided broadly into three periods. In the first period, which lasted until the early 1990s, cognitively and linguistically motivated models were tested on a few select examples. In the second, data-driven period, shallower and then statistical models were developed that could be tested on increasingly larger amounts of full texts. Finally, in the current, neural-nets-based period, the extraction of syntactic and semantic information from the text is left almost entirely to the models themselves, lexical and commonsense information is encoded using embeddings, and attentional mechanisms are incorporated in the architecture. The preneural models have been covered in great detail by Poesio et al. (2016b), so in this section we provide only a short summary of that work, and in the next section we dedicate more space to the current state of art. In Sections 4 and 6 we focus on identity reference; the other types of anaphora resolution are covered in Section 7.

4.1. Cognitively and Linguistically Rooted Early Models

The computational models proposed in the early years of research in anaphora resolution were rooted very directly in findings about anaphora from linguistic and psycholinguistic studies such as those discussed in Section 2. They focused on testing the predictions of cognitive and linguistic theories of anaphoric interpretation and therefore generally assumed a perfect syntactic and semantic analysis of the input as a starting point for anaphora resolution and/or assumed that all the needed commonsense knowledge was available.

4.1.1. Syntax-based algorithms. One of the main strands of research on computational models of anaphora focused on testing syntactic constraints and preferences on pronoun resolution. The most influential work in this area is the so-called Hobbs algorithm (Hobbs 1978), which incorporates the syntactic constraints and preferences discussed in Section 2 and provided a competitive baseline for pronoun resolution well into the data-driven period.

4.1.2. Commonsense knowledge and inference-based approaches. Much of the early work on anaphora resolution in CL (and psychology) was devoted to providing an account of the effects of commonsense knowledge and inference on the interpretation of anaphoric expressions like the one seen in example 7. The most developed proposal was the Interpretation as Abduction formal account of the inferences involved in interpreting anaphoric reference and other aspects of language interpretation, implemented in the TACITUS system (Hobbs et al. 1993). This account is possibly the most detailed one of inference in anaphora resolution, together with the less formal account by Carter (1987). But the first MUC shared tasks revealed that this approach would not scale, and thus there was a shift toward more heuristic systems in the subsequent editions of MUC.

4.1.3. Salience. The most detailed computational model of the effect of salience on the interpretation of anaphoric expressions was Sidner's (1979) focus model. In this model, focal information is used to generate salience-based preferences using very detailed (and very complex) rules, which are then assessed by commonsense inference. Two lines of research emerged from Sidner's proposals. Carter (1987) proposed a detailed model about the role of salience in interpretation and its integration with commonsense inference. A second line of work pursued simpler models of salience, leading to the development of Grosz & Sidner's (1986) theory of discourse

structure as well as Centering theory [Grosz et al. 1995 (1986), Walker et al. 1998]. Several computational models of Centering were proposed. Tetreault's Left-to-Right Centering algorithm (Tetreault 2001) was tested on a substantial data set and performed slightly better than Hobbs's algorithm. A different approach to modeling attention in anaphora resolution was the development of so-called activation-based models of salience (e.g., Lappin & Leass 1994). Such models do not hypothesize the existence of foci or centers; instead, each discourse entity has an activation level that is affected by a variety of factors.

4.1.4. Formal approaches to discourse model construction. The development of DRT and other dynamic logics led to computational models of anaphora resolution based on such theories (Alshawi 1992, Poesio 1994, Bos 2004). The best known of these models is SRI's Core Language Engine (Alshawi 1992), which is used in a number of domain-restricted practical applications. Bos's (2004) BOXER model for DRT-based semantic interpretation has been shown to be usable for large-scale semantic interpretation.

4.2. Heuristic and Knowledge-Poor Approaches

The MUC shared tasks led to a shift toward models that could be tested on a larger scale. The key characteristic of these models is that they had to perform with much more limited knowledge than the models discussed above. Unlike the early syntax-based algorithms, they could no longer assume perfect, hand-produced syntactic and semantic knowledge about the input. Instead, they had to rely on the partial or imperfect syntactic analysis produced by existing automatic parsers. And unlike knowledge-based systems, they could not expect a knowledge base that contained a complete set of axioms for all the concepts encountered. Instead, they had to rely on approximate lexical knowledge sources such as WordNet. This was the first dramatic change on the path toward fully data-driven computational models. Another change was that whereas some of these models, like most previous systems, focused on a single type of anaphoric expression, such as pronouns (Baldwin 1997, Mitkov 1998) or definite descriptions (Vieira & Poesio 2000), the systems participating in those early shared tasks had to handle all types of nominals (Kameyama 1997, Humphreys et al. 1998).

The most lasting innovation of the heuristic-based systems of this period is the precision-first architecture pioneered by COGNIAC (Baldwin 1997) (which, until recently, was still used in the Stanford Deterministic Coreference Resolver; Lee et al. 2013) and systems based on this approach. COGNIAC resolves pronouns by applying a series of rules ordered so that the most reliable apply first. The same strategy was adopted in the hand-coded version of the Viera/Poesio system (Vieira & Poesio 2000). The precision-first architecture was revived in the Stanford Deterministic Coreference Resolver for the 2011 CONLL shared task (Lee et al. 2013). The success of the Stanford Deterministic system was by all accounts due to two characteristics. First of all, the system employed a high recall and high precision component for detecting mentions. The performance of mention detection is to this day one of the most important factors in anaphora resolution. Secondly, the mentions thus extracted were processed by 10 heuristic rules, or sieves, ordered from the most accurate to the least accurate. The Stanford Sieve approach is still the best way to develop an anaphoric resolver for a language for which there are no annotated data sets.

4.3. Statistical Models of Identity Reference

4.3.1. The mention-pair model. Even if the MUC corpora and other data sets that became available at the time were fairly small, they enabled the development of the first anaphora resolution models that employed machine learning methods (Aone & Bennett 1995, Vieira & Poesio 2000).

Some of these models were essentially versions of the heuristic systems in which the optimal order among the heuristics was learned from the data, but soon more advanced models appeared—in particular, the mention-pair model proposed by Aone & Bennett (1995) and made popular by Soon et al. (2001). The mention-pair model is a simple way to recast anaphora resolution as a classification task: The model is trained to decide whether the two mentions (markables) within a pair corefer. "Resolving" potential anaphor m_j is thus viewed as the task of finding the mention m_i whose probability of coreferring with m_j is maximal: $\text{argmax}_{m_i} P(C = 1|\langle m_i, m_j \rangle)$. A coreference resolver based on this architecture

1. goes through the markables in a text (generally, but not always, in the order specified by the text);
2. for each markable m_j identifies a set of possible candidate antecedents; and
3. for each $\langle\, m_j$, candidate antecedent $m_i \,\rangle$ pair, extracts a number of features (see below) and uses them to compute the probability of $C = 1$.

Once all mentions have been classified, a clustering algorithm is used to build coreference chains out of the anaphoric links identified by the model. The Soon et al. model was the reference architecture for anaphora resolution for many years.

4.3.2. Entity-mention and mention-ranking models. From a linguistic and cognitive perspective, viewing anaphora resolution as a mention-pairing task is a drastic simplification of discourse model construction that, for instance, would appear unable to handle anaphoric references to entities not introduced via NPs. From a machine learning perspective as well, this approach would appear limited as it considers only mention and mention-pair features, not features of entities. These shortcomings led to the development of so-called entity-mention models in which mentions are directly linked to entities (clusters), as done in the prestatistical models (e.g., Luo et al. 2004, Culotta et al. 2007). A second respect in which many models have diverged from the Soon et al. (2001) architecture is the use of the "best first" approach to antecedent selection: considering multiple antecedents in parallel and choosing the one that is highest ranked instead of considering one candidate at a time. The cluster ranking model by Rahman & Ng (2011), which combined entity-mention architecture with a ranking approach, achieved state-of-the-art results for its time.

4.3.3. Extended feature sets. Another active line of research focused on improving on the Soon et al. (2001) model by employing a richer set of features. This work led to the hypothesis that richer feature sets could lead to improvements only with larger data sets than MUC. This hypothesis was indirectly confirmed by Bengtson & Roth (2008), who found that when testing on a larger data set—the ACE 2004 corpus—a simple mention-pair model using carefully chosen features could outperform the state-of-the-art system by Culotta et al. (2007). Research was also carried out on methods for mining the values of these features (Bergsma 2016).

4.3.4. Lexical and commonsense knowledge. Features that encode the lexical and commonsense knowledge required during anaphora resolution include selectional restrictions on the interpretation of pronouns (Kehler et al. 2004, Ponzetto & Strube 2006) and synonymy information and encyclopedic knowledge for interpreting nominals (Vieira & Poesio 2000, Ponzetto & Strube 2006). Another line of research focused on leveraging existing knowledge bases, such as WordNet, FrameNet, and Wikipedia (Vieira & Poesio 2000, Ponzetto & Strube 2006). One of the best-known models of this type, by Ponzetto & Strube (2006), used WordNet for lexical synonymy information, used FrameNet for selectional restrictions, and pioneered the use of Wikipedia for encyclopedic knowledge. These resources were further exploited in many models using extended

feature sets, such as those of Daume & Marcu (2005), Bengtson & Roth (2008), Rahman & Ng (2011), and Durrett & Klein (2013). In yet another line of research, distributional semantics was used to acquire semantic information from corpora (Versley et al. 2016). Several analyses of the effectiveness of lexical and commonsense knowledge for anaphora resolution (Durrett & Klein 2013, Versley et al. 2016) found the results disappointing, but as discussed below, much better results were obtained later using contextual embeddings to encode such knowledge in neural approaches.

4.3.5. Joint inference. Several interpretative tasks that affect anaphora resolution are best carried out jointly with it. One example is anaphoricity detection (Poesio & Vieira 1998, Ng & Cardie 2002a, Denis & Balridge 2007, Uryupina et al. 2016). Another example is mention detection: Daume & Marcu (2005), for instance, showed that this task, too, is best performed jointly with anaphora resolution—a finding that lies at the core of the "end-to-end" neural model currently dominating anaphora resolution and discussed in Section 6 (Lee et al. 2017). The realization that many such tasks are best performed jointly led to numerous models adopting joint inference architectures such as the Integer Linear Programming (ILP) model (Rizzolo & Roth 2016), a form of constraint programming in which constraints can be imposed on variables modeling the outcome of separate classifiers (e.g., for coreference and anaphoricity detection). The use of ILP for coreference was popularized by Denis & Balridge (2007), who applied the framework to joint anaphoricity detection and anaphora resolution, but this approach has been widely applied in anaphora resolution (Iida & Poesio 2011).

4.3.6. Graph- and tree-based architectures. Many of the most successful statistical models for the CONLL 2012 data set were based on a formulation of coreference resolution in terms of an underlying graph structure whose nodes are the mentions in a document. Two families of methods can be identified. Nicolae & Nicolae (2006) used a graph structure where the edges between the nodes encode degrees of semantic compatibility or incompatibility judgments between the mentions. A second line of research involves growing mention trees for a document, where attachment to a branch of a tree indicates coreference. Fernandes et al. (2014), who developed the top performing system at the CONLL 2012 shared task, formulated coreference resolution as the problem of recovering a latent coreference tree for a document, encoding the most likely coreference relations. Martschat & Strube (2015) argued that several popular architectures for coreference—the mention-pair model, the mention-ranking model, and the latent coreference trees model—could in fact be viewed as predicting different types of latent structures, and they developed a unified framework for training such models by using the latent structure perceptron algorithm.

4.4. Identity Anaphora in Languages Other than English

Research on zero anaphora resolution played a key role in early computational work on anaphora—for instance, in the development of Centering (Kameyama 1985). Zero anaphora resolution has remained an active area of study for Japanese because of the prevalence of zeros in the language and the availability of the NAIST corpus (Iida et al. 2007, Sasano et al. 2009). But the release of ONTONOTES spurred much research on zero pronoun anaphora in Chinese (Chen & Ng 2016) and Arabic (Aloraini & Poesio 2020) as well. A noteworthy characteristic of work on zero anaphora is that many proposals are multilingual (Iida & Poesio 2011, Aloraini & Poesio 2020); this is still sadly rare in the field notwithstanding the availability of a number of multilingual data sets.

Most topics discussed in this section are covered in greater detail by chapters in Poesio et al. (2016b). For early and heuristic models, readers are referred to Poesio et al. (2016c) (and Mitkov 2002 for more in-depth coverage). The mention-pair model with its variants is discussed in Hoste (2016). More advanced models including the entity-mention model are discussed in Ng (2016).

Bergsma (2016) and Versley et al. (2016) cover feature extraction from corpora and the use of lexical and commonsense knowledge, respectively. Joint inference is covered by Rizzolo & Roth (2016), and detection of nonanaphoricity and nonreference is discussed in Uryupina et al. (2016).

5. EVALUATION

One of the fundamental issues in anaphora resolution is that although the field has converged on an "official" metric that has driven progress for the last 10 years, it is far from clear that this metric captures our intuitions about how anaphoric interpretation should be evaluated—or indeed, what these intuitions are. This is not to say that the field is completely divided. For instance, it is universally accepted that evaluation should be entity based instead of mention based, in the sense that a system's interpretation of example 9 should be evaluated on the extent to which it recognizes that 1, 2, and 3 are all mentions of the same entity, as opposed to merely its ability to link 2 to 1 and 3 to 2.

(9) $[\text{Mary}]_i^1$ woke up late that morning, so $[\text{she}]_i^2$ rushed out of bed—$[\text{she}]_i^3$ had an important meeting.

For this reason, ever since the first MUC shared task, precision, recall, and F value for anaphora resolution have been used to assess a system's ability to identify the entire set of mentions of an entity (aka the coreference chain). However, agreement on this point still leaves many degrees of freedom. As a result, different ways have been proposed to compare the coreference chains in the gold annotation (in anaphora resolution, this is generally known as the "keys") with those produced by a system (known as "responses"), and no consensus has been reached on which metric is most appropriate. This impasse was broken by Denis & Balridge (2007), who introduced a measure based on MUC, B^3, and the Constrained Entity-Aligned F-Measure (CEAF) that was adopted in the CONLL 2011 and 2012 shared tasks and has since become standard.

5.1. A Link-Based Metric: The MUC Score

The MUC official scorer (Vilain et al. 1995) introduced a link-based metric. A link-based metric measures the extent to which the links in the response match the links in the key. For example, recall is computed by summing up the correctly recalled links for each coreference chain in the key and then dividing by the total number of correct links in the key. The number of missing links—the links found in the key entities but not in the response entities—is computed by counting the number of partitions of key K induced by response R, as follows:

$$\text{Recall}_{\text{MUC}} = \frac{\sum_i |K_i| - |\mathcal{P}(K_i; R)|}{\sum_i |K_i| - 1},$$

where $\mathcal{P}(K_i; R)$ is the partition function, which returns all the partitions of key entity K_i with respect to a system's response R.[1] Precision is computed by summing up the correct links in each coreference chain in the response and dividing that by the total number of links in the response—that is, by swapping key and response in the formula above.

5.2. A Mention-Based Metric: B^3

One problem with the MUC score is that, by definition, it only scores a system's ability to identify links between mentions; its ability to recognize that a mention does not belong to any coreference

[1]The original coreference chain gets partitioned into $k + 1$ subsets when k links are missing: One missing link results in two coreference chains, two missing links in three coreference chains, and so forth. Notice also that only $n - 1$ links are required to link all the mentions in a chain of size n.

chain—that is, its ability to classify a mention as a singleton—does not get any reward. The B^3 metric (Bagga & Baldwin 1998) was proposed to correct this problem. It does this by computing recall and precision for each mention m, even if m is a singleton. B^3 computes the intersection $|K_i \cap R_j|$ between every coreference chain K_i in the key and every coreference chain R_j in the response and then sums up recall and precision for each pair i, j and normalizes. In turn, recall and precision for i, j are computed by summing up recall and precision for each mention m in $|K_i \cap R_j|$. For instance, recall for m is the proportion of mentions in $K_i \cap R_j$ and the number of mentions in K_i:

$$\text{Recall}_{B^3}(m) = \frac{|K_i \cap R_j|}{|K_i|} \quad (m \in K_i \cap R_j).$$

Precision for m is the proportion between $|K_i \cap R_j|$ and $|R_j|$.

5.3. An Entity-Based Metric: CEAF

B^3 also suffers from a problem—namely, that a single chain in the key or response can be credited several times. This leads to anomalies; for instance, if all coreference chains in the key are merged into one in the response, the B^3 recall is one. The CEAF metric was proposed by Luo (2005) to correct this problem. The key idea of CEAF is to align chains (entities) in the key and response using a map g in such a way that each chain K_i in the key is aligned with only one chain $g(K_i)$ in the response and to then use the similarity $\phi(K_i, g(K_i))$ to compute recall and precision. Because different maps are possible, the one that achieves optimal similarity is used.

5.4. The CONLL Metric and the CONLL Scorer

After a few years in which different proposals were often difficult to compare because various researchers favored different metrics, Denis & Balridge (2007) proposed simply using the average among the F values obtained using MUC, B^3, and CEAF. This was the score used in the CONLL shared tasks in 2011 and 2012 (Pradhan et al. 2012). Since then, the reference scorer that computes this metric and takes into account a few issues that emerged (Pradhan et al. 2014) has become the standard scorer for the field (**https://github.com/conll/reference-coreference-scorers**).

5.5. The Current Practice of Evaluation in Anaphoric Reference

Although the field has now developed a unified approach to evaluation, the current practice of taking the average of three metrics cannot be considered entirely satisfactory, which is why new metrics are still being introduced every few years (for review, see Luo & Pradhan 2016, Yu et al. 2022). This aspect of current practice could certainly benefit from a reanalysis of which, if any, among the current metrics best captures linguistic intuitions or at least is best suited for practical applications (Barbu & Mitkov 2001).

There is also a need to move beyond simple identity anaphora. Because the CONLL reference scorer only scores identity reference, extended scorers were developed for the CRAC 2018 shared task (Poesio et al. 2018) and the CODI/CRAC 2021 shared task on anaphora in dialogue (Khosla et al. 2021). The scorer for the latter (Yu et al. 2022) is compatible with the old CONLL reference scorer for identity coreference but also scores the identification of singletons, nonreferring expressions, split-antecedent anaphora, bridging references, and discourse deixis and is the official scorer for the Universal Anaphora initiative (**https://github.com/juntaoy/universal-anaphora-scorer**). However, the discussion on how to evaluate these other types of anaphoric reference has only begun.

For an extensive discussion of the main evaluation metrics in coreference, and several examples explaining their working in more detail, readers are referred to Luo & Pradhan (2016); for a discussion of coreference shared tasks, readers may consult Recasens & Pradhan (2016).

6. NEURAL MODELS OF IDENTITY ANAPHORA RESOLUTION IN NEWS

6.1. Neural Networks for Coreference and the End2End Model

The paper by Wiseman et al. (2015) marked the start of the most recent shift in computational models of anaphora, from the statistical models discussed in Section 4.3 to models using neural networks to learn nonlinear functions of the input. From that point on, every improvement of the state of the art has been achieved by neural models (Lee et al. 2017, 2018; Joshi et al. 2019, 2020; Kantor & Globerson 2019; Yu et al. 2020c).

6.1.1. Embeddings. One important characteristic common to all neural models of coreference resolution is that they take as input word embeddings (Bengio et al. 2003). As discussed in the previous sections, for many years computational linguists tried to attain better lexical semantic representations by developing distributional semantics methods for learning them from corpora, but with disappointing results (Versley et al. 2016). One reason for the success of neural network models after 2010 was the emergence of a much more effective type of lexical representation, word embeddings—continuous representations learned in an unsupervised way by neural language models (Mikolov et al. 2013).

6.1.2. The End2End model. The paradigmatic neural architecture for anaphora resolution—the deep learning equivalent of the Soon et al. (2001) model—is the End2End (E2E) model proposed by Lee et al. (2017). The E2E model is a mention-pair model, but it has three key characteristics that mark a radical departure from the statistical models discussed in Section 4.3. First, as the name suggests, mention detection and antecedent identification are carried out jointly. The advantages of carrying out these tasks jointly had already been demonstrated by Daume & Marcu (2005); Lee et al. and subsequent researchers such as Yu et al. (2020a) provided conclusive evidence that with OntoNotes-size data sets, carrying out these two tasks jointly is the optimal solution. Like all neural models for anaphora resolution, the E2E model takes as input a sequence of word embeddings x_i instead of a simple bag of words. The model considers all possible spans of these words and computes a span representation—a candidate mention representation—for each. Pairs of these span representations form the mention pairs classified by the model. The second important characteristic of the model is the span representation itself, how it is computed, and the notion of "headedness" it uses. A neural network—a bidirectional LSTM—is used to compute word representations $x^*_{s,i}$ for each word $x_{s,i}$ in span s, and an attention layer is then used to assign a relative weight $\alpha_{s,i}$ to each word, out of which a weighted representation $\hat{x}_s$ is then computed for the whole span. The span representation g_s is then specified as a quadruple $g_s = [x^*_{s,\text{START}}, x^*_{s,\text{END}}, \hat{x}_s, \phi_s]$ consisting of the word representations for the first and last word in the span, the weighted representation, and a couple of other features. This means that the model learns, and in a task-specific way, (*a*) what is the best representation for the mention as a whole, and (*b*) a "soft" notion of head assigning a weight to each of the words in the NP, including modifiers, determiners, and so forth. This approach is believed to address many of the difficulties identified in earlier work (e.g., how to define heads in a general way) and is one of the key reasons for the success of the E2E model.

6.1.3. Learned features. The third crucial feature of Lee et al.'s (2017) model is that it takes only word embeddings as features. This is another clear difference from the statistical models

discussed in Section 4.3. Like all modern deep learning models, Lee et al.'s E2E model is able to learn by itself almost all the linguistic generalizations required by anaphora resolution directly from the data, without the kind of feature engineering required even by statistical models.

6.1.4. Performance. What made the E2E model the reference architecture for anaphora resolution is that it achieved a 68.8 CONLL score (see Section 5), which was 6 points higher than the state-of-the-art statistical model (Martschat & Strube 2015). And as discussed below, this performance could be further improved by adopting more advanced embeddings.

6.1.5. Cluster ranking. Another direction of research aimed at improving the E2E model focuses on using cluster ranking instead of mention ranking (Lee et al. 2018, Kantor & Globerson 2019, Yu et al. 2020c). For instance, the entity equalization model by Kantor & Globerson (2019) resulted in significant improvements through their method for building cluster representations out of mention representations. The cluster ranking model by Yu et al. (2020c) also achieved significant improvements and is notable as the only model discussed in this section to also carry out nonreferring expressions and singleton identification.

6.2. From Static to Context-Sensitive Embeddings

Shortly after the E2E model was proposed, another major technical innovation in deep learning resulted in further and substantial improvements: the development of so-called context-sensitive embeddings like ELMO (Peters et al. 2018) and BERT (Devlin et al. 2019). These are embeddings that, unlike earlier pretrained embeddings such as Word2Vec (Mikolov et al. 2013), assign different interpretations to words depending on the context. For many years CL researchers had tried without success to demonstrate that wordsense disambiguation in context was important (Versley et al. 2016); ELMO and BERT provided conclusive evidence for this. Adding ELMO to the E2E model immediately resulted in an improvement of more than six percentage points over the original version of the model, from 68.6 to 73.0 on CONLL score (Lee et al. 2018). The subsequent development of the BERT model (Devlin et al. 2019) resulted in another three-percentage-point improvement (Joshi et al. 2019, Kantor & Globerson 2019). More recently still, the SPANBERT approach to training BERT with the type of spans used in anaphora resulted in another three-percentage-point improvement on ONTONOTES. That is, the performance on ONTONOTES has improved by almost 20 percentage points in the space of 5 years. Crucially, a big part of this improvement is due to the fact that these models supply much of the knowledge required by an anaphoric interpreter.

6.3. The State of the Art for Identity Reference in News

The current state of the art for anaphora resolution in news articles from ONTONOTES and from ARRAU, in which nonreferring expressions and singletons are also annotated, is summarized in **Table 2**. The table reports the results on ONTONOTES of (the latest version of) the two highest-performing statistical models in the CONLL 2012 shared task (Björkelund & Kuhn 2014, Fernandes et al. 2014) as well as the results of two statistical models that further pushed performance on that data set, followed by the best-known neural models prior to the E2E model, and then by the models using increasingly more sophisticated context-sensitive embeddings. We also provide the results on ARRAU of the only neural model (Yu et al. 2020c) that reported results on that data set.

7. BEYOND IDENTITY ANAPHORA IN NEWS

Virtually all the research discussed in Section 6 focuses on the resolution of identity reference in news text from the ONTONOTES corpus. However, one of the most exciting developments of the last 5 years is that the field is moving beyond this narrow focus to cover anaphora resolution in

Table 2 State of the art on news

Year	Model	MUC (F_1)	B^3 (F_1)	CEAF (F_1)	CONLL
ONTONOTES					
2014	Fernandes et al. 2014	70.5	57.6	53.9	60.6
	Björkelund & Kuhn 2014	70.7	58.6	55.6	61.6
	Durrett & Klein 2014	71.2	58.7	55.2	61.7
2015	Martschat & Strube 2015	72.2	59.6	55.7	62.5
2015	Wiseman et al. 2015	72.6	60.5	57.1	63.4
2017	End2End (Lee et al. 2017)	77.2	66.6	62.6	68.8
2018	Lee et al. 2018	80.4	70.8	67.6	73.0
2019	Kantor & Globerson 2019	83.4	74.7	71.8	76.6
	Joshi et al. 2019	83.5	75.3	71.9	76.9
2020	Yu et al. 2020c	83.0	74.7	71.6	76.4
	Joshi et al. 2020	85.3	78.1	75.3	79.6
ARRAU (CRAC 2018 data)					
2020	Yu et al. 2020c	78.2	78.8	76.8	77.9

other genres such as scientific documents or fiction, addressing the challenges raised by benchmark data sets such as the Winograd Schema Challenge, and looking at other types of anaphora, such as bridging reference and discourse deixis.

7.1. Other Genres: Scientific Documents, Dialogue

7.1.1. Biomedical texts. After news, the most researched genre in anaphora resolution is scientific articles—in particular, in the biomedical domain. Data sets such as GENIA (Yang et al. 2004) and CRAFT (Cohen et al. 2017) have supported the development of several models for this genre and a number of shared tasks, the best known of which are those organized in connection with the BIONLP workshops (Nguyen et al. 2011, Baumgartner et al. 2019). More recently, this genre has witnessed the deployment of systems using embeddings specially trained for scientific and biomedical texts (Zhang et al. 2019).

7.1.2. Dialogue and conversational agents. Although some of the initial work on anaphora in CL was motivated by research on question-answering systems and task-oriented dialogue systems (e.g., Webber 1979), most research in the data-driven period has been focused on written text, for lack of suitable corpora. The few exceptions typically have involved the researchers creating the necessary data sets themselves (Poesio 1994, Byron 2002, Müller 2008). The one language for which substantial corpora of anaphora in dialogue exist is French: The ANCOR corpus (Muzerelle et al. 2014) has enabled the development of the end-to-end neural model for coreference interpretation such as that of Grobol (2020). It is hoped that the data sets created in the recent shared tasks on Anaphora Resolution in Dialogue (Khosla et al. 2021) will encourage more research in this genre.

7.2. The Winograd Schema Challenge

Computational work on the Winograd Schema Challenge can be categorized in a broadly similar way to research on anaphora resolution in news. The first computational models were statistical models such as that of Rahman & Ng (2012). More recently, however, most models for this task have been neural. The top performing among such systems use pretrained language models such

as BERT (Devlin et al. 2019)—indeed, such models are often assessed using benchmarks including the Winograd Schema Challenge as a subtask, such as GLUE (Wang et al. 2019). An example of a system using BERT is that of Kocijan et al. (2019).

7.3. Bridging Reference Resolution

Bridging reference resolution is a popular area of research because it involves modeling both inference and salience, two of the most studied preferences in anaphoric interpretation (Sidner 1979). In work following Sidner's, the emphasis shifted to how to acquire the required knowledge, whether from lexical resources (Vieira & Poesio 2000) or from corpora (Poesio et al. 2004a, Markert & Nissim 2005). Also, whereas early work on bridging resolution mostly focused on bridging reference via definite nominals (Sidner 1979, Vieira & Poesio 2000), later systems covered all types of bridging references (Poesio et al. 2004a, Hou et al. 2018, Roesiger et al. 2018, Yu & Poesio 2020). The work of Hou et al. (2018) represents the current state of the art on full bridging resolution, but it was evaluated only on ISNOTES. The first neural model for bridging reference resolution was proposed by Yu & Poesio (2020).

7.4. Other Types of Anaphora

7.4.1. Discourse deixis. There has not been a lot of work on discourse deixis resolution. The first implemented anaphora resolution system resolving discourse deixis is Byron's (2002) PHORA rule-based algorithm for pronoun interpretation in dialogue. The first machine learning–based models for pronoun resolution covering both anaphora and discourse deixis were proposed by Müller (2008). Kolhatkar et al. (2013) concentrated on resolving definite nominals containing what they called shell nouns: nouns like *issue*, which have a preferential abstract interpretation. A key innovation from Kolhatkar et al. (2013) was the use of large amounts of synthetically created training data to alleviate data sparsity, which is the main problem with using machine learning to develop models for discourse deixis. This approach to data creation, which was further developed by Marasović et al. (2017), is receiving increasing attention for rare aspects of anaphora resolution.

7.4.2. Split-antecedent plurals. Early research on split-antecedent anaphora (Eschenbach et al. 1989, Kamp & Reyle 1993) mostly focused on the constraints on the construction of complex entities from singular entities. More recent studies, such as that of Vala et al. (2016), focused on a subset of the problem. The first neural system for resolving split-antecedent anaphora that are expressed by both pronouns and other types of NPs was developed by Yu et al. (2020b), testing on ARRAU. Yu et al. (2021) proposed the first neural system resolving both single and split-antecedent anaphora and not requiring gold mention input.

Cohen et al. (2017) provide a useful survey not only of existing data sets for coreference resolution in biomedical texts but also of proposed models for the genre. The literature on tackling the Winograd Schema Challenge is systematically surveyed by Kocijan et al. (2020). The recent article by Kobayashi & Ng (2020) reviews the literature on bridging, whereas the literature on discourse deixis is systematically covered by Kolhatkar et al. (2018).

SUMMARY POINTS

1. It is an exciting time to work on anaphora resolution. The field now has data sets and benchmarks of sufficient size and quality to support large-scale investigation of at least identity anaphora, in multiple languages.

2. We also have evaluation methods that appear to be effective at assessing the relative quality of systems, at least for identity anaphora.
3. These data sets and evaluation methods have spurred rapid progress in the field; models like Lee et al.'s (2017) End2End, in combination with context-sensitive embeddings, have addressed many of the problems that appeared unsolvable just 10 years ago.

FUTURE ISSUES

1. To ensure that we have a genuine understanding of anaphoric interpretation, we need to branch out in terms of data sets beyond written news and scientific text, to cover at least language use in spoken conversations and in narrative.
2. We also need to start paying more attention to other aspects of anaphoric interpretation that are very common, such as discourse deixis.
3. Even if we restrict our attention to identity anaphora, important differences remain between the existing schemes, and addressing these will require addressing the types of hard questions about anaphoric annotation raised by Zaenen (2006) (e.g., Which types of context dependence should an anaphoric resolver attempt to interpret? Can this question be answered in isolation from an application?).
4. We also need to start addressing the issue of disagreement and subjectivity in anaphoric interpretation and its implications for the field.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was funded in part by the DALI project (ERC grant 695662) and in part by the ARCIDUCA project (EPSRC grant EP/W001632/1).

LITERATURE CITED

Aloraini A, Poesio M. 2020. Cross-lingual zero pronoun resolution. In *Proceedings of the 12th International Conference on Language Resources and Evaluation* (*LREC 2020*), pp. 90–98. Paris: Eur. Lang. Resour. Assoc.

Alshawi H, ed. 1992. *The Core Language Engine*. Cambridge, MA: MIT Press

Aone C, Bennett SW. 1995. *Automatic acquisition of anaphora resolution strategies*. Paper presented at the AAAI Spring Symposium on Empirical Methods in Discourse: Interpretation and Generation, Mar. 27–29, Palo Alto, CA

Asher N, Lascarides A. 2003. *The Logic of Conversation*. Cambridge, UK: Cambridge Univ. Press

Bagga A, Baldwin B. 1998. Algorithms for scoring coreference chains. In *Proceedings of the LREC Workshop on Linguistic Coreference*, pp. 563–66. Paris: Eur. Lang. Resour. Assoc.

Baldwin B. 1997. CogNIAC: high precision pronoun coreference with limited knowledge and precision preferences. In *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, ed. R Mitkov, B Boguraev, pp. 38–45. Stroudsburg, PA: Assoc. Comput. Linguist.

Bamman D, Lewke O, Mansoor A. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the 12th International Conference on Language Resources and Evaluation* (*LREC 2020*), pp. 44–54. Paris: Eur. Lang. Resour. Assoc.

Barbu C, Mitkov R. 2001. Evaluation tool for rule-based anaphora resolution methods. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 34–41. Stroudsburg, PA: Assoc. Comput. Linguist.

Baumgartner WA Jr., Bada M, Pyysalo S, Ciosici MR, Hailu N, et al. 2019. Craft shared tasks overview—integrated structure, semantics, and coreference. In *Proceedings of the 5th Workshop on BioNLP*, pp. 174–84. Stroudsburg, PA: Assoc. Comput. Linguist.

Bengio Y, Ducharme R, Vincent P, Jauvin C. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.* 3:1137–55

Bengtson E, Roth D. 2008. Understanding the value of features for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 294–303. Stroudsburg, PA: Assoc. Comput. Linguist.

Bergsma S. 2016. Extracting anaphoric agreement properties from corpora. See Poesio et al. 2016b, pp. 345–68

Björkelund A, Kuhn J. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47–57. Stroudsburg, PA: Assoc. Comput. Linguist.

Bos J. 2004. Computational semantics in discourse: underspecification, resolution, and inference. *J. Logic Lang. Inform.* 13(2):139–57

Büring D. 2005. *Binding Theory*. Cambridge, UK: Cambridge Univ. Press

Byron DK. 2002. Resolving pronominal references to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 80–87. Stroudsburg, PA: Assoc. Comput. Linguist.

Carter DM. 1987. *Interpreting Anaphors in Natural Language Texts*. Chichester, UK: Ellis Horwood

Chen C, Ng V. 2016. Chinese zero pronoun resolution with deep neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 778–88. Stroudsburg, PA: Assoc. Comput. Linguist.

Chinchor NA, Sundheim B. 1995. *Message Understanding Conference (MUC) tests of discourse processing*. Paper presented at the AAAI Spring Symposium on Empirical Methods in Discourse: Interpretation and Generation, Mar. 27–29, Palo Alto, CA

Clark HH. 1977. Bridging. In *Thinking: Readings in Cognitive Science*, ed. PN Johnson-Laird, P Wason, pp. 411–20. London/New York: Cambridge Univ. Press

Cohen KB, Lanfranchi A, Choi MJ-y, Bada M, Baumgartner WA Jr., et al. 2017. Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. *BMC Bioinform.* 18:372

Culotta A, Wick M, McCallum A. 2007. First-order probabilistic models for coreference resolution. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 81–88. Stroudsburg, PA: Assoc. Comput. Linguist.

Daume H III, Marcu D. 2005. A large-scale exploration of effective global features for a joint entity detection and tracking model. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 97–104. Stroudsburg, PA: Assoc. Comput. Linguist.

Denis P, Baldridge J. 2007. 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*, pp. 236–43. Stroudsburg, PA: Assoc. Comput. Linguist.

Devlin J, Chang M-W, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–86. Stroudsburg, PA: Assoc. Comput. Linguist.

Doddington G, Mitchell A, Przybocki M, Ramshaw L, Strassell S, Weischedel R. 2000. The Automatic Content Extraction (ACE) program—tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (*LREC 2004*), pp. 837–40. Paris: Eur. Lang. Resour. Assoc.

Durrett G, Klein D. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1971–82. Stroudsburg, PA: Assoc. Comput. Linguist.

Durrett G, Klein D. 2014. A joint model for entity analysis: coreference, typing, and linking. *Trans. Assoc. Comput. Linguist.* 2:477–90

Eschenbach C, Habel C, Herweg M, Rehkämper K. 1989. Remarks on plural anaphora. In *EACL '89: Proceedings of the Fourth Conference on European Chapter of the Association for Computational Linguistics*, pp. 161–67. Stroudsburg, PA: Assoc. Comput. Linguist.

Fernandes ER, dos Santos CN, Milidiú RL. 2014. Latent trees for coreference resolution. *Comput. Linguist.* 40(4):801–35

Fox BA. 1987. *Discourse Structure and Anaphora*. Cambridge, UK: Cambridge Univ. Press

Garnham A. 2001. *Mental Models and the Interpretation of Anaphora*. New York: Psychol. Press

Grobol L. 2020. *Coreference resolution for spoken French*. PhD Thesis, Univ. Sorbonne Nouv., Paris

Grosz BJ. 1977. *The representation and use of focus in dialogue understanding*. PhD Thesis, Stanford Univ., Stanford, CA

Grosz BJ, Joshi AK, Weinstein S. 1995 (1986). Centering: a framework for modeling the local coherence of discourse. *Comput. Linguist.* 21(2):202–25

Grosz BJ, Sidner CL. 1986. Attention, intention, and the structure of discourse. *Comput. Linguist.* 12(3):175–204

Gundel JK, Abbott B, eds. 2019. *The Oxford Handbook of Reference*. Oxford, UK: Oxford Univ. Press

Heim I. 1982. *The semantics of definite and indefinite noun phrases*. PhD Thesis, Univ. Mass., Amherst

Hendrickx I, Bouma G, Coppens F, Daelemans W, Hoste V, et al. 2008. A coreference corpus and resolution system for Dutch. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (*LREC 2008*), pp. 144–49. Paris: Eur. Lang. Resour. Assoc.

Hinrichs EW, Kübler S, Naumann K. 2005. A unified representation for morphological, syntactic, semantic and referential annotations. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pp. 13–20. Stroudsburg, PA: Assoc. Comput. Linguist.

Hobbs JR. 1978. Resolving pronoun references. *Lingua* 44:311–38

Hobbs JR, Stickel M, Appelt D, Martin P. 1993. Interpretation as abduction. *Artif. Intell. J.* 63:69–142

Hoste V. 2016. The mention-pair model. See Poesio et al. 2016b, pp. 269–82. Berlin: Springer

Hou Y, Markert K, Strube M. 2013. 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*, pp. 907–17. Stroudsburg, PA: Assoc. Comput. Linguist.

Hou Y, Markert K, Strube M. 2018. Unrestricted bridging resolution. *Comput. Linguist.* 44(2):237–84

Humphreys K, Gaizauskas R, Azzam S, Huyck C, Mitchell B, et al. 1998. *University of Sheffield: description of the LaSIE-II System as used for MUC-7*. Paper presented at the Seventh Message Understanding Conference (MUC-7), Fairfax, VA, Apr. 29–May 1

Ide N, Pustejovsky J, eds. 2017. *The Handbook of Linguistic Annotation*. Dordrecht, Neth.: Springer

Iida R, Inui K, Matsumoto Y. 2007. Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Trans. Asian Lang. Inf. Process.* 6(4):1

Iida R, Poesio M. 2011. A cross-lingual ILP solution to zero anaphora resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 804–13. Stroudsburg, PA: Assoc. Comput. Linguist.

Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O. 2020. SpanBERT: improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguist.* 8:64–77

Joshi M, Levy O, Zettlemoyer L, Weld D. 2019. BERT for coreference resolution: baselines and analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5803–8. Stroudsburg, PA: Assoc. Comput. Linguist.

Kameyama M. 1985. *Zero anaphora: the case of Japanese*. PhD Thesis, Stanford Univ., Stanford, CA

Kameyama M. 1997. Recognizing referential links: an information extraction perspective. In *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pp. 46–53. Stroudsburg, PA: Assoc. Comput. Linguist.

Kamp H, Reyle U. 1993. *From Discourse to Logic*. Dordrecht, Neth.: D. Reidel

Kamp H, von Genabith J, Reyle U. 2011. Discourse representation theory. In *Handbook of Philosophical Logic*, ed. D Gabbay, F Guenthner, pp. 125–394. Dordrecht, Neth.: Springer

Kantor B, Globerson A. 2019. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 673–77. Stroudsburg, PA: Assoc. Comput. Linguist.

Karttunen L. 1976. Discourse referents. In *Syntax and Semantics 7 - Notes from the Linguistic Underground*, ed. J McCawley, pp. 363–85. New York: Academic

Kehler A, Appelt D, Taylor L, Simma A. 2004. The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 289–96. Stroudsburg, PA: Assoc. Comput. Linguist.

Khosla S, Yu J, Manuvinakurike R, Ng V, Poesio M, et al. 2021. The CODI-CRAC 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pp. 1–15. Stroudsburg, PA: Assoc. Comput. Linguist.

Kobayashi H, Ng V. 2020. Bridging resolution: a survey of the state of the art. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3708–21. n.p.: Int. Comm. Comput. Linguist.

Kocijan V, Cretu AM, Camburu OM, Yordanov Y, Lukasiewicz T. 2019. A surprisingly robust trick for the Winograd Schema Challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4837–42. Stroudsburg, PA: Assoc. Comput. Linguist.

Kocijan V, Lukasiewicz T, Davis E, Marcus G, Morgenstern L. 2020. A review of Winograd Schema Challenge datasets and approaches. arxiv:2004.13831 [cs.CL]

Kolhatkar V, Roussel A, Dipper S, Zinsmeister H. 2018. Anaphora with non-nominal antecedents in computational linguistics: a survey. *Comput. Linguist.* 44(3):547–612

Kolhatkar V, Zinsmeister H, Hirst G. 2013. Interpreting anaphoric shell nouns using antecedents of cataphoric shell nouns as training data. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 300–10. Stroudsburg, PA: Assoc. Comput. Linguist.

Landragin F. 2016. Description, modéelisation et détection automatique des chaînes de réeférence (DEMOCRAT). *Bull. Assoc. Fr. Intell. Artif.* 92:11–15

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

Lee H, Chang A, Peirsman Y, Chambers N, Surdeanu M, Jurafsky D. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Comput. Linguist.* 39(4):885–916

Lee K, He L, Lewis M, Zettlemoyer LS. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 188–97. Stroudsburg, PA: Assoc. Comput. Linguist.

Lee K, He L, Zettlemoyer LS. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 687–92. Stroudsburg, PA: Assoc. Comput. Linguist.

Levesque HJ, Davis E, Morgenstern L. 2012. The Winograd Schema Challenge. In *KR'12: Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, ed. G Brewka, T Eiter, SA McIlraith, pp. 552–61. Palo Alto, CA: AAAI Press

Luo X. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 25–32. Stroudsburg, PA: Assoc. Comput. Linguist.

Luo X, Ittycheriah A, Jing H, Kambhatla N, Roukos S. 2004. A mention-synchronous coreference resolution algorithm based on the Bell tree. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (*ACL-04*), pp. 135–42. Stroudsburg, PA: Assoc. Comput. Linguist.

Luo X, Pradhan S. 2016. Evaluation metrics. See Poesio et al. 2016b, pp. 141–63

Lyons J. 1977. *Semantics*. Cambridge, UK: Cambridge Univ. Press

Mann WC, Thompson SA. 1988. Rhetorical Structure Theory: toward a functional theory of text organization. *Text* 8(3):243–81

Marasović A, Born L, Opitz J, Frank A. 2017. A mention-ranking model for abstract anaphora resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 221–32. Stroudsburg, PA: Assoc. Comput. Linguist.

Markert K, Hou Y, Strube M. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 795–804. Stroudsburg, PA: Assoc. Comput. Linguist.

Markert K, Nissim M. 2005. Comparing knowledge sources for nominal anaphora resolution. *Comput. Linguist.* 31(3):367–402

Martschat S, Strube M. 2015. Latent structures for coreference resolution. *Trans. Assoc. Comput. Linguist.* 3:405–18

McEnery A, Tanaka I, Botley S. 1997. Corpus annotation and reference resolution. In *ANARESOLUTION '97: Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, ed. R Mitkov, B Boguraev, pp. 67–74. Stroudsburg, PA: Assoc. Comput. Linguist.

Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems*, Vol. 2, pp. 3111–19. Red Hook, NY: Curran

Mitkov R. 1998. 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*, Vol. 2, pp. 869–75. Stroudsburg, PA: Assoc. Comput. Linguist.

Mitkov R. 2002. *Anaphora Resolution*. London: Longman

Müller MC. 2008. *Fully automatic resolution of* it, this *and* that *in unrestricted multy-party dialog*. PhD Thesis, Univ. Tübingen, Tübingen, Ger.

Muzerelle J, Lefeuvre A, Schang E, Antoine JY, Pelletier A, et al. 2014. Ancor_centre, a large free spoken French coreference corpus: description of the resource and reliability measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (*LREC 2014*), pp. 843–47. Paris: Eur. Lang. Resour. Assoc.

Nedoluzhko A, Mírovský J, Ocelák R, Pergler J. 2009. *Extended coreferential relations and bridging anaphora in the Prague Dependency Treebank*. Presented at the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), Goa, India, Nov. 5–6

Nedoluzhko A, Novák M, Popel M, Žabokrtský Z, Zeman D. 2021. *Coreference meets universal dependencies—a pilot experiment on harmonizing coreference datasets for 11 languages*. ÚFAL Tech. Rep. TR-2021-66, Charles Univ., Prague

Ng V. 2016. Advanced machine learning models for coreference resolution. See Poesio et al. 2016b, pp. 283–313

Ng V, Cardie C. 2002a. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *COLING 2002: The 19th International Conference on Computational Linguistics*. n.p.: Int. Comm. Comput. Linguist. **https://aclanthology.org/C02-1139/**

Ng V, Cardie C. 2002b. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics*, pp. 104–11. Stroudsburg, PA: Assoc. Comput. Linguist.

Nguyen NLT, Kim JD, Tsujii J. 2011. Overview of the protein coreference task in BioNLP shared task. In *Proceedings of the BioNLP Shared Task Workshop*, pp. 74–82. Stroudsburg, PA: Assoc. Comput. Linguist.

Nicolae C, Nicolae G. 2006. BESTCUT: a graph algorithm for coreference resolution. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 275–83. Stroudsburg, PA: Assoc. Comput. Linguist.

Ogrodniczuk M, Głowińska K, Kopeć M, Savary A, Zawisławska M. 2015. *Coreference in Polish: Annotation, Resolution and Evaluation*. Berlin: Walter de Gruyter

Partee BH. 1972. Opacity, coreference, and pronouns. In *Semantics for Natural Language*, ed. D Davidson, G Harman, pp. 415–41. Dordrecht, Neth.: D. Reidel

Passonneau RJ. 1997. *Instructions for applying discourse reference annotation for multiple applications (DRAMA)*. Work. Pap., Columbia Univ., New York

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, et al. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, *Volume 1 (Long Papers)*, pp. 2227–37. Stroudsburg, PA: Assoc. Comput. Linguist.

Poesio M. 1994. *Discourse interpretation and the scope of operators*. PhD Thesis, Univ. Rochester, Rochester, NY

Poesio M. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the Workshop on Discourse Annotation*, pp. 72–79. Stroudsburg, PA: Assoc. Comput. Linguist.

Poesio M. 2016. Linguistic and cognitive evidence about anaphora. See Poesio et al. 2016b, pp. 23–54

Poesio M, Artstein R. 2005. The reliability of anaphoric annotation, reconsidered: taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, ed. A Meyers, pp. 76–83. Stroudsburg, PA: Assoc. Comput. Linguist.

Poesio M, Artstein R. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (*LREC 2008*), pp. 1170–74. Paris: Eur. Lang. Resour. Assoc.

Poesio M, Bruneseaux F, Romary L. 1999. The MATE meta-scheme for coreference in dialogues in multiple languages. In *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging*, ed. M Walker, pp. 65–74. Stroudsburg, PA: Assoc. Comput. Linguist.

Poesio M, Chamberlain J, Kruschwitz U, Paun S, Uma A, Yu J. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *Volume 1 (Long and Short Papers)*, pp. 1778–89. Stroudsburg, PA: Assoc. Comput. Linguist.

Poesio M, Grishina Y, Kolhatkar V, Moosavi N, Roesiger I, et al. 2018. Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference*, *Anaphora and Coreference*, pp. 11–22. Stroudsburg, PA: Assoc. Comput. Linguist.

Poesio M, Mehta R, Maroudas A, Hitzeman J. 2004a. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (*ACL-04*), pp. 143–50. Stroudsburg, PA: Assoc. Comput. Linguist.

Poesio M, Pradhan S, Recasens M, Rodriguez K, Versley Y. 2016a. Annotated corpora and annotation tools. See Poesio et al. 2016b, pp. 97–140

Poesio M, Stevenson R, Di Eugenio B, Hitzeman JM. 2004b. Centering: a parametric theory and its instantiations. *Comput. Linguist.* 30(3):309–63

Poesio M, Stuckardt R, Versley Y. 2016b. *Anaphora Resolution: Algorithms*, *Resources and Applications*. Berlin: Springer

Poesio M, Stuckardt R, Versley Y, Vieira R. 2016c. Early approaches to anaphora resolution: theoretically inspired and heuristic-based. See Poesio et al. 2016b, pp. 55–94

Poesio M, Vieira R. 1998. A corpus-based investigation of definite description use. *Comput. Linguist.* 24(2):183–216

Ponzetto S, Strube M. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the NAACL*, *Main Conference*, pp. 192–99. Stroudsburg, PA: Assoc. Comput. Linguist.

Pradhan S, Luo X, Recasens M, Hovy E, Ng V, Strube M. 2014. Scoring coreference partitions of predicted mentions: a reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 30–35. Stroudsburg, PA: Assoc. Comput. Linguist.

Pradhan S, Moschitti A, Xue N, Uryupina O, Zhang Y. 2012. CoNLL-2012 shared task: modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Joint Conference on EMNLP and CoNLL - Shared Task*, pp. 1–40. Stroudsburg, PA: Assoc. Comput. Linguist.

Rahman A, Ng V. 2011. Narrowing the modeling gap: a cluster-ranking approach to coreference resolution. *J. Artif. Intell. Res.* 40:469–521

Rahman A, Ng V. 2012. Resolving complex cases of definite pronouns: the Winograd Schema Challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 777–89. Stroudsburg, PA: Assoc. Comput. Linguist.

Recasens M, Hovy E, Martí MA. 2011. Identity, non-identity, and near-identity: addressing the complexity of coreference. *Lingua* 121(6):1138–52

Recasens M, Màrquez L, Sapena E, Martí MA, Taulé M, et al. 2010. SemEval-2010 Task 1: coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation* (*SemEval 2010*), pp. 1–8. Stroudsburg, PA: Assoc. Comput. Linguist.

Recasens M, Martí MA. 2010. AnCora-CO: coreferentially annotated corpora for Spanish and Catalan. *Lang. Resourc. Eval.* 44(4):315–45

Recasens M, Pradhan S. 2016. Evaluation campaigns. See Poesio et al. 2016b, pp. 165–208

Rizzolo N, Roth D. 2016. Integer linear programming for coreference resolution. See Poesio et al. 2016b, pp. 315–43

Roesiger I, Riester A, Kuhn J. 2018. Bridging resolution: task definition, corpus resources and rule-based experiments. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3516–28. Stroudsburg, PA: Assoc. Comput. Linguist.

Sakaguchi K, Le Bras R, Bhagavatula C, Choi Y. 2020. WINOGRANDE: an adversarial Winograd Schema Challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pp. 8732–34. Palo Alto, CA: AAAI Press

Sanford AJ, Garrod SC. 1981. *Understanding Written Language*. Chichester, UK: Wiley

Sasano R, Kawahara D, Kurohashi S. 2009. The effect of corpus size on case frame acquisition for discourse analysis. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 521–29. Stroudsburg, PA: Assoc. Comput. Linguist.

Sidner CL. 1979. *Towards a computational theory of definite anaphora comprehension in English discourse*. PhD Thesis, MIT, Cambridge, MA

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

Tetreault JR. 2001. A corpus-based evaluation of Centering and pronoun resolution. *Comput. Linguist.* 27(4):507–20

Uryupina O, Artstein R, Bristot A, Cavicchio F, Delogu F, et al. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *J. Nat. Lang. Eng.* 26(1):95–128

Uryupina O, Kabadjov MA, Poesio M. 2016. Detecting non-reference and non-anaphoricity. See Poesio et al. 2016b, pp. 369–92

Vala H, Piper A, Ruths D. 2016. 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)*, pp. 2287–96. Stroudsburg, PA: Assoc. Comput. Linguist.

van Deemter K, Kibble R. 2000. On coreferring: coreference in MUC and related annotation schemes. *Comput. Linguist.* 26(4):629–37

Versley Y, Poesio M, Ponzetto SP. 2016. Using lexical and encyclopedic knowledge. See Poesio et al. 2016b, pp. 393–429

Vieira R, Poesio M. 2000. An empirically-based system for processing definite descriptions. *Comput. Linguist.* 26(4):539–93

Vilain M, Burger J, Aberdeen J, Connolly D, Hirschman L. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference* (*MUC-6*), pp. 45–52. San Francisco: Morgan Kaufmann

Walker MA, Joshi AK, Prince EF, eds. 1998. *Centering Theory in Discourse*. Oxford, UK: Clarendon

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S. 2019. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–55. Stroudsburg, PA: Assoc. Comput. Linguist.

Webber BL. 1979. *A Formal Approach to Discourse Anaphora*. New York: Garland

Webber BL. 1991. Structure and ostension in the interpretation of discourse deixis. *Lang. Cogn. Process.* 6(2):107–35

Webster K, Recasens M, Axelrod V, Baldridge J. 2018. Mind the GAP: a balanced corpus of gendered ambiguous pronouns. *Trans. Assoc. Comput. Linguist.* 6:605–17

Winograd T. 1972. Understanding natural language. *Cogn. Psychol.* 3(1):1–191

Wiseman SJ, Rush AM, Shieber SM, Weston J. 2015. 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)*, pp. 1416–26. Stroudsburg, PA: Assoc. Comput. Linguist.

Xia P, Durme BV. 2021. Moving on from OntoNotes: coreference resolution model transfer. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 5241–56. Stroudsburg, PA: Assoc. Comput. Linguist.

Yang X, Zhou G, Su J, Tan CL. 2004. Improving noun phrase coreference resolution by matching strings. In *Lecture Notes in Computer Science*, Vol. 3248: *Natural Language Processing—IJCNLP 2004*, ed. KY Su, J Tsujii, JH Lee, OY Kwong, pp. 22–31. Berlin: Springer

Yu J, Bohnet B, Poesio M. 2020a. Neural mention detection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation* (*LREC 2020*), pp. 1–10. Paris: Eur. Lang. Resour. Assoc.

Yu J, Khosla S, Moosavi N, Paun S, Pradhan S, Poesio M. 2022. The Universal Anaphora scorer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (*LREC 2022*), pp. 4873–83. Paris: Eur. Lang. Resour. Assoc.

Yu J, Moosavi N, Paun S, Poesio M. 2021. Stay together: a system for single and split-antecedent anaphora resolution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4174–84. Stroudsburg, PA: Assoc. Comput. Linguist.

Yu J, Moosavi NS, Paun S, Poesio M. 2020b. Free the plural: unrestricted split-antecedent anaphora resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6113–25. n.p.: Int. Comm. Comput. Linguist.

Yu J, Poesio M. 2020. Multitask learning-based neural bridging reference resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3534–46. n.p.: Int. Comm. Comput. Linguist.

Yu J, Uma A, Poesio M. 2020c. A cluster ranking model for full anaphora resolution. In *Proceedings of the 12th International Conference on Language Resources and Evaluation* (*LREC 2020*), pp. 11–20. Paris: Eur. Lang. Resour. Assoc.

Zaenen A. 2006. Mark-up barking up the wrong tree. *Comput. Linguist.* 32(4):577–80

Zeldes A. 2020. *Multilayer Corpus Studies*. New York/London: Routledge

Zeldes A. 2022. Can we fix the scope for coreference? *Dialogue Discourse* 13(1):41–62

Zhang H, Song Y, Song Y, Yu D. 2019. Knowledge-aware pronoun coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 867–76. Stroudsburg, PA: Assoc. Comput. Linguist.