information

Article

Coreference Resolution: Toward End-to-End and Cross-Lingual Systems

André Ferreira Cruz , Gil Rocha * and Henrique Lopes Cardoso

LIACC/DEI, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal; andre.ferreira.cruz@fe.up.pt (A.F.C.); hlc@fe.up.pt (H.L.C.)

* Correspondence: gil.rocha@fe.up.pt

Received: 30 December 2019; Accepted: 28 January 2020; Published: 30 January 2020

Abstract: The task of coreference resolution has attracted considerable attention in the literature due to its importance in deep language understanding and its potential as a subtask in a variety of complex natural language processing problems. In this study, we outlined the field's terminology, describe existing metrics, their differences and shortcomings, as well as the available corpora and external resources. We analyzed existing state-of-the-art models and approaches, and reviewed recent advances and trends in the field, namely end-to-end systems that jointly model different subtasks of coreference resolution, and cross-lingual systems that aim to overcome the challenges of less-resourced languages. Finally, we discussed the main challenges and open issues faced by coreference resolution systems.

Keywords: natural language processing; coreference resolution; end-to-end systems; cross-language learning

1. Introduction

Natural language processing (NLP) is a scientific field concerned with endowing computing machines with the ability to process, interpret, or generate natural (i.e., human) language, both in spoken and written forms. Considering the latter, several tasks have been addressed by the NLP community, ranging from syntactic to semantic analysis of text, including parsing and discourse analysis. Assessing the semantic meaning of text is particularly demanding, and several specific challenges have been addressed by the research community. With this study, we addressed one such challenge.

Coreference resolution is an NLP task that involves determining all referring expressions that point to the same real-world entity. A referring expression (i.e., a mention) is either a noun phrase (NP), a named entity (NE), or a pronoun, which refer to an entity in the real world known as the referent [1]. A grouping of referring expressions with the same referent is called a coreference chain or cluster. The goal of a coreference resolution system is to output all the coreference chains of a given text. In the literature, the term partial entity is often used interchangeably with coreference chain.

With its roots in the 1960s, this core NLP task is still far from being solved, as it often relies on common sense reasoning, which is difficult for machines to acquire. Figure 1 provides three separate examples and their corresponding coreference chains. A classification algorithm could, for instance, use the hyponym/hypernym semantic relationship between "bee" and "insect" to classify the two mentions as co-referent, and use world-knowledge to detect a strong relationship between the mentions "Barack Obama" and "president".

1. [Bees]$_0$ are critical to safeguarding [food supplies worldwide]$_1$. [These interesting insects]$_0$ have been hit hard by [climate change]$_2$.
2. [Barack Obama]$_0$, [the former US president]$_0$, has told [the country]$_1$ [he]$_0$'s ready for [a long vacation]$_2$.
3a. [The city councilmen]$_0$ refused [the demonstrators]$_1$ [a permit]$_2$ because [they]$_1$ advocated violence.
3b. [The city councilmen]$_0$ refused [the demonstrators]$_1$ [a permit]$_2$ because [they]$_0$ feared violence.

Figure 1. Coreference resolution examples. The third example was extracted from the Winograd Schema Challenge [2].

Addressing this problem typically requires extensive previous language processing, such as named-entity recognition, part of speech tagging, parsing, and semantic analysis. In several works, a pipeline system to fetch useful word relationships from world knowledge sources is also used [1,3,4]. This variety of subtasks and knowledge sources, most of which are also emerging topics in the literature, attests to the difficulty of coreference resolution.

The limitations machines exhibit when performing coreference resolution are evidenced by the use of this task in tests of machine intelligence. The Winograd Schema Challenge [2] is a multiple-choice test that was designed to be an improvement on the Turing test. The multiple choice questions included in the challenge rely on the resolution of a type of anaphora, a noun phrase followed by a coreferent pronoun, in contexts where pronoun disambiguation requires understanding the meaning of further elements in the sentence. One such question is shown in the third example in Figure 1, in which the referent of the "they" pronoun can switch from "the demonstrators" (3a) to "the city councilmen" (3b) by simply changing the subsequent word from "advocated" to "feared".

Ng [5] conducted an extensive study on the task of entity coreference resolution, but since then, several novel systems have been proposed, extending the state-of-the-art technology on the standard OntoNotes dataset benchmark by more than 11 F1 points [6]. This one-year improvement is larger than achieved in the previous four years combined. Ng did not address the issue of biases in the task, the available resources, related tasks, or the topic of cross-lingual coreference resolution.

The problem of coreference resolution is closely related to that of anaphora resolution [7]. Despite anaphora being potentially seen as a type of coreference, this is not an accurate definition. As observed by Sukthanker et al. [8], anaphora resolution is essentially intralinguistic in the sense that the entities are present in the text, and resolving them does not typically require world knowledge. Coreference resolution, conversely, concerns terms with potentially different meanings, yet referring to the same extra-linguistic world entity.

In this study, we explored the current state-of-the-art advances in the field of coreference resolution and their interwoven relationship with common NLP tasks. We list reference resources used for training state-of-the-art systems on different languages and for introducing world knowledge at different levels of the coreference resolution pipeline. Our aim was to provide an extensive introduction for machine learning practitioners that are newcomers to the field of coreference resolution, containing essential information for expediting state-of-the-art contributions. This work also serves as a lightly opinionated reference for active researchers in the area to review current trends, future directions, and interactions with other research fields.

The rest of the paper is organized as follows. In Section 2, we introduce several NLP tasks that are closely related to coreference resolution, and show how they have been approached in connection with the latter. In Section 3, we outline available corpora as well as available external resources for world knowledge extraction. In Section 4, we explore the most-used coreference resolution metrics, plus their differences and shortcomings. In Section 5, we delve into the state-of-the-art approaches to coreference resolution together with their limitations. In Section 6, we examine some recent trends in NLP in general, and in coreference resolution in particular, including neural, end-to-end, and cross-lingual approaches. In Section 7, we outline the broader challenges for this task, from biases to common sense reasoning. Finally, in Section 8 we provide some final considerations on the directions of coreference resolution research.

2. Related Tasks

Coreference resolution typically requires a pre-processing pipeline comprising a variety of NLP tasks (e.g., tokenization, lemmatization, named entity recognition, part-of-speech tagging). Historically, these tasks are addressed before training the coreference resolution model (in a pre-processing stage) and, consequently, errors made by pre-processing models impact coreference resolution models, which typically assume that the information provided from this pre-processing stage is correct. With less-resourced languages, trained models on these pre-processing tasks are typically less accurate due to the lack of annotated data for each task. As a result, the cascade of errors resulting from the commonly used pipeline approach is more impactful on these resource-scarce languages. Conversely, NLP systems requiring deep language understanding (e.g., quotation attribution, textual entailment, and argument mining) generally benefit from coreference information.

Several promising approaches have been reported to address coreference resolution jointly with other NLP tasks, showing improved results on both counts, such as Almeida et al. [9], Hajishirzi et al. [4], and Durrett and Klein [10]. In this section, we present some of these tasks and explore their relationship with coreference resolution.

2.1. Named Entity Recognition

Named entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities in text into predefined categories such as the names of persons, organizations or locations [11]. Identifying the token-level boundaries of mentions (such as named entities) is a necessary step toward obtaining mention clusters from natural text; thus, NER is tightly related to coreference resolution. Additionally, categorizing entity mentions in text provides useful insight to a coreference system. It is beneficial to know whether "Apple" or "Amazon" refer to a brand or their nature counterparts when assessing coreferences, since such concepts have disparate lexical and semantic relationships. Named entity disambiguation (NED) is a task devoted to this purpose.

Coreference resolution and named entity recognition were first jointly tackled by Daumé III and Marcu [12], who represented the possible mention partitions as nodes in a tree (a node at the *i*th level corresponds to a partition of the first *i*th mentions) [5]. The goal is then to search for the most likely leaf node, allowing for the exploitation of cluster-level features computed over the coreference chains created so far. This system performs both tasks in a non-pipelined way, enabling the mention detection module to take advantage of coreference information, and vice versa.

Lee et al. [13] developed a neural model that considers all spans of text up to a maximum length of 10 words as possible mentions, and uses the same internal representations of these spans to simultaneously classify them as an entity mention (or not) and as corefering (or not) with another text span. This type of end-to-end system has received increasing attention in years, and end-to-end joint tackling of NER and coreference resolution in has been featured in most state-of-the-art systems (see Section 6.2).

2.2. Entity Linking

Named entity linking (NEL), or "Wikification", is the task of resolving named entity mentions to entries in a structured knowledge base NEL is useful wherever it is necessary to compute with direct reference to people, places, and organizations, rather than potentially ambiguous or redundant character strings [4]. Besides disambiguating a mention (as is completed in NED), NEL provides deeper knowledge of the real-world entity referred in a text, providing key insight towards the possible coreferencing mentions (e.g., "Apple" is a "tech company"). Several models have been proposed to tackle these tasks jointly [4,10], showing clear performance improvements for both tasks.

Recently, several shared tasks oriented toward jointly tackling mention detection and entity linking to a knowledge base, with particular focus on a cross-lingual setting, were presented, showing an increasing interest in exploring cross-lingual techniques to boost the performance of NLP system across different languages [14–17].

2.3. Part-of-Speech Tagging

Part-of-speech (PoS) tagging is the process of automatic assigning of parts of speech to words in a sentence [18]. State-of-the-art systems for this core NLP task has surpassed 97% accuracy since 2003 [19,20]. The most recent advances in this task pushed the state-of-the-art performance to 97.96% accuracy [21] on the standard Wall Street Journal portion of the Penn Treebank dataset [22].

Regarding coreference resolution, PoS tagging is often used as part of a pipeline system for preprocessing data, as it is a valuable feature for classifying mentions as coreferent or not [1,4,23]. Gender and number match are known indicators of coreference affinity, and the empirically measured benefits of these and other features were detailed by Bengtson and Roth [24].

2.4. Quotation Attribution

The problem of quotation attribution consists of automatically attributing quotes to speakers [24]. Reported speech attribution has far-reaching uses, from fact checking (through finding corroborating news stories and articles) to opinion mining. Speaker information is extensively used as a feature in coreference resolution, from older systems [25] to recent state-of-the-art models [26], as it is known to contain useful information for this task.

Joint models have also been used to improve the performance of quotation attribution. For instance, Almeida et al. [9] modeled the problem as finding an optimal spanning tree in a graph containing both quotation nodes and mention nodes.

3. Resources

In this section, we revise available annotated corpora to address coreference resolution, as well as the resources commonly used in this task for semantic tagging, mention encoding (i.e., feature representation of the mentions) and entity linking, among other related tasks. Additionally, we discuss corpora scarcity in languages other than English and its impact on the progression of research in this area.

3.1. Corpora

When using supervised machine learning techniques, as customary in the state-of-the-art systems for coreference resolution (see Section 5), the availability of annotated corpora is an important requirement. Large-scale corpora have been built for the English language, the most prominent being the OntoNotes 5.0 dataset [27]. For other languages, available corpora are typically smaller in size. This scarcity poses a considerable barrier to improving coreference resolution of low-resource languages, which may be tackled using unsupervised approaches (see Section 5.5) or transfer learning from higher-resourced languages [28,29], a technique that is becoming more frequent.

Corpora available for this task often differ on the annotation scheme used, the domain from which they were extracted, and the type of labeled coreferences. Concerning the latter, Message Understanding Conference (MUC) corpora [30] do not annotate singleton mentions (non-corefering); Automatic Content Extraction (ACE) [31] features singleton mentions, but is restricted to seven semantic types (person, organization, geo-political entity, location, facility, vehicle, and weapon); and corpora from SemEval-2010 [32], as well as OntoNotes [27], feature coreference relationships between all types of noun phrases. Regarding the annotation format, both the SemEval-2010 and OntoNotes corpora follow CoNLL's [27], as do most general domain corpora released after the 2012 CoNLL shared task.

A compilation of the most prominent corpora for coreference resolution that include the English language is featured in Table 1 in chronological order, including their size in number of tokens and documents. These general domain corpora were selected due to their size (larger than 100,000 tokens for English) and/or historical significance (MUC datasets).

Most corpora are extracted from a variety of day-to-day sources (e.g., newswire, broadcast, blogs, and Wikipedia), leading to similar resolution class distributions and resulting in domain-agnostic corpora. However, coreference resolution systems are known to degrade performance when transferred across dissimilar domains, such as legal texts or scientific papers. Chaimongkol et al. [33] showed that although the several MUC and ACE corpora have strong correlations with the distribution of resolution classes among them; a corpus of annotated scientific papers has uncorrelated resolution class distributions (–0.10 correlation coefficient) against the sum of general domain corpora.

As such, several corpora were developed for domain-specific coreference resolution. For the scientific domain, the corpus from Schäfer et al. [34] (1.3 million tokens) is composed of papers from ACL anthology, Chaimongkol et al. [33] (42,000 tokens) contains scientific papers from several research areas, and Cohen et al. [35] (28,000 tokens) is comprised of biomedical journal articles.

Other corpora focus on a cross-lingual setting, resulting from shared tasks on the topic. The TAC-KBP (Knowledge Base Population track at the Text Analytics Conference) shared tasks [15–17] have used cross-lingual entity discovery and linking corpora since 2015. These feature data in English, Chinese, and Spanish, with the addition of 10 low-resource languages in the 2017 corpora. The latest 2017 TAC-KBP corpus [17] features 500 documents and 231,000 tokens of gold-standard coreference data, evenly distributed between English, Chinese, and Spanish, whereas the remaining 10 languages have varying data sizes, from silver- or gold-standard sources. The multi-lingual nature of these corpora allow an entity discovery and linking (EDL) system to track entities along documents in different languages, which, in addition to the entity linking phase, have increased complexity over regular coreference resolution.

In the same direction, the CORBON 2017 shared task [36] presented a large parallel corpora with silver-standard coreference annotations, aimed at the development of projection-based approaches to coreference resolution. More recently, Nedoluzhko et al. [37] presented a multi-lingual parallel treebank with gold-standard coreference annotations, consisting of English text translated into Czech, Russian, and Polish.

Several other corpora exist for this task, which are not described here due to their smaller size or their specialized research setting. We refer the reader to Sukthanker et al. [8] for an extensive compilation of the currently available corpora.

Table 1. Collection of general domain coreference resolution corpora.

Corpus	Language	# Tokens	# Documents
MUC-6 [38]	English	25,000	60
MUC-7 [30]	English	40,000	67
ACE (2000-2004) [31]	English	960,000	-
	Chinese	615,000	-
	Arabic	500,000	-
SemEval-2010 [32]	English	120,000	353
	Catalan	345,000	1138
	Dutch	104,000	240
	German	455,000	1235
	Italian	140,000	143
	Spanish	380,000	1183
OntoNotes v5.0 [27]	English	1,600,000	2384
	Chinese	950,000	1729
	Arabic	300,000	447

3.2. External Semantic Resources

Several available semantic resources can be exploited to aid in the coreference resolution task. A common example is the use of lexical databases, which store the lexical category and synonyms of words, as well as semantic relationships between word pairs or triplets (e.g., hyponym/hypernym). These include the Princeton WordNet for English [39], a large-scale lexical database organized around the notion of a synset, which is a set of words with the same part-of-speech that can be used interchangeably in particular contexts. Following this project, several other lexical databases have been developed, namely the EuroWordNet [40], a multilingual wordnet comprising several inter-linked wordnets for European languages. There are other types of lexical databases, a good example of which is BabelNet [41], a sizeable multilingual semantic network featuring lexicographic and encyclopedic knowledge.

With the accelerating popularity of deep learning, recent works have indicated an increasing usage of word embeddings [13,42–44]: distributed mappings of words to dense feature vectors. If the dataset used to train the embeddings is sufficiently large, these distributed representations will inherently encode semantic similarities and world-knowledge relationships [45–47]. However, as is increasingly problematic in machine learning, this type of statistical world knowledge may hide systemic biases [48]; this issue is discussed in Section 7.1.

The combination of word embeddings and deep neural networks are the basis of most recent state-of-the-art NLP systems. The use of multi-lingual word embeddings [49] enables the training of language-agnostic systems, although language transfer is still far from perfect—see Section 6.3.

Details on how these external semantic resources have been explored to tackle coreference resolution are presented in Section 7.3.

4. Evaluation Metrics

Machine learning classification tasks are typically evaluated using precision, recall, and F-scores. For coreference resolution, however, a variety of task-specific metrics have been designed to better translate a system's scores into real-world performance. In this section, we overview the most widely used metrics for evaluating coreference resolution systems, pointing out the main characteristics of each metric from a practical and high-level perspective. Further details were beyond the scope of this study. We refer the reader to Moosavi and Strube [50] for a detailed explanation of each metric.

Dating back to the Sixth Message Understanding Conference, the MUC scoring algorithm is a link-based evaluation scheme that operates by comparing the equivalence classes defined by the gold standard links and the classifier's links [51]; that is, it determines how many links would need to be added to obtain the correct clustering. Although commonly used, this metric has some shortcomings: firstly, as it is link-oriented, it disregards singleton mentions; secondly, its link-based F-measure inherently benefits systems producing fewer mentions [52].

To address MUC's shortcomings, Bagga and Baldwin [53] proposed the B^3 scoring algorithm. This metric computes individual precision and recall for each mention by looking at its presence/absence in the final mention clusters, and then computes a weighted sum of these values. However, this metric assumes the system's mention set to be the gold standard, resulting in problematic system mentions that are not mapped to any gold-standard mention, and vice versa—an issue referred to as twinless mentions by Stoyanov et al. [54]. A few variants of this algorithm have since been introduced to overcome this problem, namely by Stoyanov et al. [54] and Rahman and Ng [55].

Luo [52] constructed the constrained entity-alignment F-Measure (CEAF), claiming to overcome MUC's pitfalls and to impvoe interpretability. The CEAF metric computes the alignment between gold and system entities; it measures the similarity of each mention cluster (representing an entity) to determine the value of each possible alignment. The best alignment is then used to calculate the CEAF precision, recall, and F-measure. Cai and Strube [56] introduced two variants of the B^3 and CEAF metrics to overcome the issue of different gold and system mention sets, enabling the evaluation of systems whose input is natural text (without gold standard mention boundaries).

More recently, Recasens and Hovy [57] proposed the BLANC measure, criticizing the B^3 and CEAF metrics for inflating a system's score by overly rewarding singleton mentions. This flaw is easily visible in baseline systems that classify all mentions as singletons. In the English portion of the SemEval-2010 corpora, this baseline achieved scores of 71.2 CEAF F1 and 83.2 B^3 F1, but only 49.6 points on the BLANC measure [32]. The best performing system on this shared task [58] achieved scores of 74.3 CEAF F1, 60.8 MUC F1, 82.4 B^3 F1, and 70.8 BLANC, very close to the baseline on the CEAF and B^3 measures, but distinctively better on the BLANC metric.

The BLANC measure adapts the Rand index, a performance metric for clustering algorithms [59] for coreference resolution. The original version of this metric also assumed identical sets of true mentions and system mentions [57], with Luo et al. [60] later extending the metric's definition to overcome this problem.

More recently, Moosavi and Strube [50] identified an underlying barrier to these metrics' interpretability, the mention identification effect: the more mentions a system identifies, regardless of whether they are correct, the more accurate the system's performance. Moosavi and Strube's system showed decreased performance on CEAF, B^3, and BLANC metrics when artificially removing all incorrectly resolved mentions entirely from the response entities, and increased performance when artificially adding incorrectly linked mentions that had not been identified (mentions that existed in the key set but not in the response set). The only metric that is resistant to the mention identification effect is MUC, which is also the least discriminative metric, and has other known issues (as discussed above). As a result, the same authors proposed a link-based entity aware (LEA) system. LEA considers the size of an entity as a measure of its importance and evaluates resolved coreference relations instead of resolved mentions, claiming to overcome the identified shortcomings.

This considerable diversity of metrics imposes a question regarding how to accurately compare system performance. Although an active research topic, virtually all coreference resolution systems developed since 2011 adopt the CoNLL metric: the unweighted average of MUC, B^3, and CEAF scores [27].

5. State-of-the Art Models

In this section, we outline the main machine learning approaches to coreference resolution. We start by summarizing the common ground of the different systems, and follow with in-depth subsections for each model type. A model may fit in multiple categories; in such cases, we place it in its most defining one. Table 2 reports the results of several systems that have driven improvements in the state-of-the-art systems in recent years, benchmarked on the English portion of the OntoNotes dataset.

The common architecture of a coreference resolution system features a data preparation phase and a resolution phase as seen in Figure 2. The data preparation pipeline consists of the detection of mentions in the input text (NER), followed by a feature extraction step, converting each data instance into an expressive feature vector. The resolution phase consists of the classification of these instances as coreferent or not, and the linking of mentions and partial entities into the final coreference chains. These two steps of the resolution phase can run simultaneously or separately. If run separately, the linking step is performed on top of the output produced by the classification step (i.e., based on the pair-wise predictions of coreferent mentions, the linking step will determine the final coreference chains).

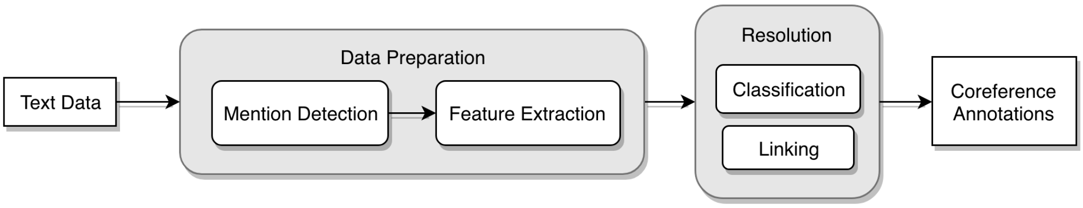

Figure 2. Typical architecture of a coreference resolution system.

Table 2. Results on the English portion of the data from CoNLL-2012 shared tasks in decreasing order of average F1 score (CoNLL metric). The bottom model is the only unsupervised model with the rest being supervised. Only non-ensembled scores are listed.

	MUC	B^3	$CEAF_{\phi 4}$	**CoNLL**
	F1	**F1**	**F1**	**(Avg. F1)**
Joshi et al. [6]	83.5	75.3	71.9	76.9
Kantor and Globerson [61]	83.4	74.7	71.2	76.6
Fei et al. [84]	81.4	71.7	68.4	73.8
Lee et al. [26]	80.4	70.8	67.6	73.0
Peters et al. [43]	78.6	68.1	64.6	70.4
Zhang et al. [85]	76.5	65.5	61.4	67.8
Lee et al. [13]	75.8	65.0	60.8	67.2
Clark and Manning [42]	74.6	63.4	59.2	65.7
Clark and Manning [74]	74.2	63.0	58.7	65.3
Wiseman et al. [73]	73.4	61.5	57.7	64.2
Wiseman et al.l [72]	72.6	60.5	57.1	63.4
Clark and Manning [66]	72.6	60.4	56.0	63.0
Martschat and Strube [86]	72.2	59.6	55.7	62.5
Durrett and Klein [10]	71.2	58.7	55.2	61.7
Björkelund and Kuhn [87]	70.7	58.6	55.6	61.6
Durrett and Klein [25]	69.2	57.5	54.3	60.3
Ma et al. [83]	67.7	55.9	51.8	58.4

However, a different and increasingly popular approach relies on end-to-end differentiable systems, having recently achieved the best performance on the OntoNotes dataset benchmark [6,13,26,61]. We discuss this type of system and other new trends in Section 6.

For clarification, we use the term *end-to-end system* to refer to a system that performs coreference resolution from natural text in a non-pipelined manner. As such, these systems jointly model mention detection and coreference resolution (and potentially other related tasks), thus better preventing error cascading. This term is sometimes used in the literature to refer to all systems that perform coreference resolution from natural text, both pipelined and non-pipelined systems alike [56]; however, we do not adopt this terminology. Instead, we follow that of several other recent papers [13,26,62].

Coreference resolution systems can be further subdivided according to how they handle the resolution phase, regarding how the classification and linking steps operate. Several approaches perform classification and linking as two separate steps, enabling the use of global optimization techniques in the linking phase, such as path-finding [63], clustering [64], or graph-partitioning algorithms [1]. Additionally, some approaches use heuristics to select the best antecedent mention from a pool of positively identified antecedents [5]. In contrast, some systems execute classification and linking in one step, performing both tasks online and simultaneously [1,65,66].

In the following subsections, we examine the major learning-based models for coreference resolution: the well-studied mention-pair, entity-mention, mention-ranking, and cluster-ranking models, which differ in the different approaches followed in each phase of the coreference resolution system described above.

5.1. Mention-Pair Models

Mention-pair models determine whether two mentions are coreferent by classifying the specific mention-pair (often with a confidence value). The classifier generates a feature vector using local information to encode the mentions and their relationships. The mention-pair predictions completed by the classifier are later used in the linking process to obtain the final coreference chains. Despite their apparent simplicity, these are arguably the most popular type of coreference classifiers, with consistently published works [1,54,67,68].

The training data for this model are typically generated by creating a set of mention-pair learning instances for every mention combination in the training document. This creates an unbalanced dataset with a high ratio of negative to positive examples [69]. We discuss this frequent challenge and its possible solutions in Section 7.

The mention-pair model has two major weaknesses: lack of global information in the classification phase, considering mentions independently of each other, and expressiveness limitations as the features extracted from a single mention-pair may not be sufficient to properly classify it, especially in the case of pronouns. Additionally, the gender of a noun or pronoun is often undefined, thus aggravating the need for other non-local sources of information, as seen in Figure 3. This problem is particularly common in Germanic languages (e.g., English, German, and Swedish), as opposed to the Romance languages (e.g., Spanish, Italian, Portuguese, French), which do not feature neutral pronouns.

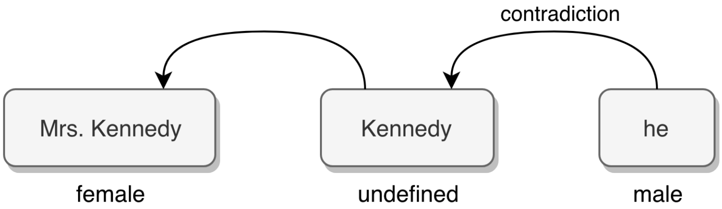

Figure 3. Example of contradictions in the linking process [1]. Arrows represent positive coreference links.

Regarding the linking process, Klenner and Ailloud [64] used a clustering algorithm with good results, Finkel et al. [70] considered it as a global optimization problem, Yang et al. [23] tackled it with a more frequent heuristic-based approach, and Luo et al. [63] intified the problem as finding the best path from the root of a Bell tree to the leaf nodes.

5.2. Mention Rankers

Instead of considering the coreference resolution problem as a pairwise binary classification task between all possible antecedents, mention-ranking models select the best antecedent for a given target mention from a group of ranked candidate antecedents, possibly selecting no antecedent (a special NA mention). This approach limits the errors caused by the lack of contextual global information, but may still encounter contradictions during the classification phase by considering each target mention and the corresponding anaphora separately, which must later be solved in the linking process. By ranking mentions instead of imposing their coreferring link, this type of models provides better global optimization linking strategies. For instance, Denis and Baldridge's [71] approach delivered considerable improvements over the state-of-the-art methods by minimizing a global ranking loss function.

Representing a step up from the simpler mention-pair model, this type of model is consistently found in the literature [25,72]. The state-of-the-art end-to-end system by Lee et al. [13] adopts a similar span-ranking approach, considering all spans of text as possible mentions and then ranking their pairwise coreference affinity (see Section 6.2).

5.3. Entity-Mention Models

Mention-pair models have limitations in their expressiveness, as they can only apply features local to the mentions being classified at the moment. To overcome this issue, researchers investigated the development of entity-mention coreference models [1,63,65]. These models classify a partial entity and a mention, or two partial entities, as coreferring or not.

Entity-based classifiers generally have the same features as mention-pair classifiers, the difference being that an entity feature's value considers the specific values of all mentions in a partial entity.

Consequently, this type of model is often able to overcome contradictions in the linking process. For example, a gender-neutral name such as "Kennedy" may provoke linking contradictions if featured alongside the mentions "Mrs. Kennedy" and "he". As seen in Figure 3, a mention-pair model may link the mentions "Mrs. Kennedy" and "Kennedy" on the basis of a partial string match, followed by nearby mentions "Kennedy" and "he" and lack of gender information in the first mention. In contrast, an entity-mention model should be able to overcome this contradiction (between "Mrs. Kennedy" and "he") by considering the entity's gender (female) instead of the mention's gender (undefined).

5.4. Cluster Rankers

Cluster-ranking models take a pair of mention clusters (partial entities) instead of a pair of mentions as input. These partial entities are incrementally built of coreferring mentions, starting from a cluster of a single mention and iteratively joining coreferential clusters. This type of model is therefore able to overcome local linking contradictions using entity-level information.

Wiseman et al. [73] used recurrent neural networks (RNNs) to extract entity-level representations of mentions, reevaluating a mention's representation on each new link by running a RNN over the cluster's mentions in order. This was found to be particularly useful for pronominal mentions, as these do not encode enough information to be independently resolved.

Clark and Manning [74] further improved the performance of cluster-ranking models using a pooling-based cluster-pair encoder. Each mention from cluster $c1$ pairs with each mention of cluster $c2$, the pairs are run through a mention-pair encoder, and the resulting matrix of mention-pair representations is max- and average-pooled for a final cluster-pair representation.

Although more recent mention-ranking systems have surpassed the performance of cluster-ranking approaches, providing global entity-level information is mandatory for a complete coreference resolution system. Information is lacking in most pronominal mention-pairs to be able to consistently solve them. Humans do not consider each mention independently, instead requiring the context of the whole document to resolve most coreferences.

5.5. Unsupervised and Semi-Supervised Models

As customary in supervised learning tasks, the need for large gold-standard corpora is a bottleneck to the advancement of research. In addition, these corpora are highly expensive to create and are thus scarce and not available in low-resourced languages, or in specific domains.

Unsupervised learning approaches find and examine patterns in unlabeled data, grouping data points by different measures of similarity. The ability to learn from raw data prevents the need for expensive annotated datasets, although some labeled data are always necessary for evaluation purposes. Semi-supervised models are often bootstrapped from a small amount of labeled data, building up a larger training dataset by iteratively predicting the labels for unlabeled data and training the model on the new training data.

Gasperin [75] experimented with semi-supervised models in the context of biomedical texts, aiming to overcome the lack of in-domain corpora for the task of coreference resolution by bootstrapping resolution models from a small amount of annotated data. More recently, Raghavan et al. [76] tackled this task on the similar domain of clinical text using co-training and multi-view learning [77] with posterior regularization. The authors achieved results comparable with supervised models by co-training a pair of max entropy classifiers on extracted semantic and temporal features.

Unsupervised models aim to probabilistically infer coreference partitions on unlabeled documents. As such, several works in the literature focus on fine-tuning probabilistic generative models for this task [68,78,79]. Haghighi and Klein [78] presented a generative non-parametric Bayesian model based on a hierarchical Dirichlet process [80]. The authors used a cluster-based mixture model with linguistic features such as head-word and salience. The basic idea of head-word is that for each phrase or text-span, there is a specific word of particular significance [81] (e.g., "[Hundreds of people] fled [their homes]" [31]). Soon after, Poon and Domingos [79] presented an unsupervised system competitive with the state-of-the-art models by performing joint inference across mentions and using Markov logic to express relations and constraints between mentions. To this goal, Ng [68] recast the task as an expectation maximization (EM) clustering process [82], enabling global-level optimization and achieving encouraging results through a generative and unsupervised model.

Whereas this type of models has been explored for several years, it has consistently lagged behind supervised models performance-wise, and this gap widened after the release of the OntoNotes dataset. The best performing unsupervised coreference resolution system to date was constructed by Ma et al. [83], shown at the bottom of Table 2.

6. Current Trends

As with most other machine learning areas, coreference resolution has seen several novel approaches to the task in recent years. Consequently, state-of-the-art performance has been improving at a fast pace: the best-performing system to date [6] outperforms the best-performing system from the previous year [26] by 3.9 CoNLL points, which in turn outperformed the previous year's best-performer [13] by 5.8 CoNLL points. In this section, we focus on some recent trends: the proliferation of neural models, the direction toward end-to-end systems, and cross-lingual approaches.

6.1. Neural Models

A trend in the field (as in many other NLP tasks) is the use of neural networks to learn non-linear models of coreference resolution. Wiseman et al. [72] used this approach to learn non-linear representations of raw features in an attempt to automatically learn more useful intermediate representations. As seen in the previous section, Clark and Manning [74] used a neural cluster-ranking model for extracting entity-level representations. This model learns high-dimensional vector representations of pairs of coreference clusters, culminating in substantial improvements over the state-of-the-art model on the OntoNotes dataset. The same authors later applied deep reinforcement learning [42] to this task, directly optimizing coreference evaluation metrics instead of the typical heuristic loss function, producing further improved results.

Recent works have also shown that using ensemble neural models, trained with different random initializations, and averaging their coreference/antecedent scores in run-time yield consistent performance gains [88], although being computationally expensive. Lee et al. [13] reported a 1.6 F1 improvement when using ensembles, and Zhang et al. [85] reported 1.4 F1 improvement. These steady performance gains have led to recent works often publishing single-model performance as well as ensemble-performance for fairer comparison.

6.2. End-to-End Models

End-to-end models jointly tackle the sub-tasks of coreference resolution in a single model, allowing for related sub-tasks (i.e., mention-detection and linking, as well as other auxiliary tasks such as head-finding) to share one optimization goal, and therefore produce better overall results compared with pipelined systems. This type of modelling allows the system to learn from pairwise interactions between tasks, thus preventing cascading errors and providing more information to tasks that would traditionally occur in the beginning of the pipeline and would not have access to results from later-occurring tasks.

An end-to-end system for tackling mention detection and coreference resolution jointly was first proposed by Daumé III and Marcu [12]. Their search-based system predicts the coreference links of a document in a left-to-right manner, incorporating increasingly global features. Singh et al. [89] used joint inference of entities, their relationships, and coreference links through an adapted algorithm for belief propagation. They reported improved performance on all tasks versus a system that independently models each task.

Durrett and Klein [10] also tackled this problem by jointly modelling coreference resolution and named-entity recognition, with the addition of modelling entity linking. Their system uses a novel approach in maintaining uncertainty about all inference decisions instead of using greedy approximations. The authors used the OntoNotes and ACE datasets for coreference and entity inference, and matched entities to Wikipedia entries. Although the OntoNotes dataset does not provide gold-standard entity links, the model still uses information in Wikipedia links to improve coreference and named entity decisions (e.g., the linking of the entity "Dell" to a company's Wikipedia page may help coreference decisions in "[Dell]$_0$, founded by [Michael Dell]$_1$, . . . "). The joint modelling of these tasks leads to an improvement of 1.4 F1 for NER and 0.5 on the CoNLL metric for coreference resolution, both on the OntoNotes dataset (3.3 and 0.2 improvements on the ACE dataset, respectively). This supports the conclusion that the first task in a pipelined system has the largest potential for improvements by being jointly modelled with tasks occurring later in the pipeline, as information flows unidirectionally in a traditional NLP pipeline from earlier tasks to later tasks.

Recently, Lee et al. [13] devised a neural end-to-end differentiable system that jointly learns which spans of text are mentions and how to best link them. The authors produced vector embeddings of text spans by means of a bidirectional long short-term memory (LSTM) network [90], capturing the spans' representations in the context of the full sentence. The aforementioned work uses a head-finding attention mechanism [91], as the head word is known to be a valuable feature [24], reducing reliance on syntactic parsers and other external resources. The antecedent and mention scores for each span-pair are computed from the same learned span representations, allowing for the error signal of both tasks to be used for one common optimization goal. As all spans (up to a maximum number of tokens) are considered as potential mentions, this model is computationally intensive. However, only local information was used when ranking span-pairs, leaving room for improvement.

Peters et al. [43] later developed contextualized word embeddings, which, when built upon this span-ranking architecture from Lee et al. [13], further extended the coreference resolution (improvement from 67.2 F1 to 70.4 F1 points). Zhang et al. [85] also improved this span-ranking architecture using a biaffine attention mechanism to better capture head word information when computing antecedent scores. The biaffine attention enables the cluster scoring function to directly model both the compatibility of the two mentions (spans) and the prior likelihood of a link between them [92]. The authors incorporated both mention scoring and antecedent scoring in the loss function, optimizing both simultaneously. Lee et al. [13] optimized for cluster performance, acting as indirect supervision for mention detection.

Subsequently, Lee et al. [26] improved on their own system by adapting their span-ranking architecture to better capture entity-level information. The authors used an attention mechanism over previously predicted coarse clusters, enabling the model to iteratively refine the predicted clusters and their representations. Additionally, the system uses an antecedent pruning mechanism for reducing computational complexity, computing less expensive (and less accurate) antecedent and cluster scores for longer-distance span links (the authors used fine-grained scores for only 50 antecedents per span). This approach led to significant reported improvements on the OntoNotes dataset (73.0 on the CoNLL metric). Fei et al. [84] further improved upon this model by adapting it to a reinforcement learning setting, thus directly optimizing coreference evaluation metrics instead of a heuristic loss function, achieving 74.1 on the CoNLL metric.

Kantor and Globerson [61] tackled the problem of capturing the properties of an entire mention cluster when representing a single mention by means of an entity equalization mechanism. This method represents each mention in a cluster as an approximation of the sum of all mentions in the cluster while maintaining end-to-end differentiability. Coupled with BERT [44] embeddings, this approach pushed the leading performance to 76.6 on the CoNLL metric. Finally, Joshi et al. [6] explored the improvements produced when using the model by Lee et al. [26] with BERT embeddings, both for base and large variants. The authors found that using BERT-large provides improvements of 3.9% F1 points on the CoNLL metric when compared with the original ELMo-based model [43], achieving 76.9 on the CoNLL metric. This corresponds to the best reported performance on the English portion of the CoNLL-2012 task, as seen in Table 2. When facing BERT's 512 limit for context tokens, using independent BERT instances on non-overlapping segments proved to be the best-performing option, opposed to using overlapping segments to provide the model with context beyond 512 tokens (76.9 F1 vs. 76.1 F1, respectively). The use of BERT embeddings and the more general transformer architecture [44] have revolutionized several fields of natural language processing [93–95], advancing the limits of the field as a whole.

End-to-end systems are now commonplace among the NLP literature. They both contribute to boosting performance on virtually all tasks and promote task-independent intermediate representations, presenting a seemingly fruitful research direction. This type of system pairs nicely with learned word embeddings, which are known to be multitask learners as well [47].

6.3. Cross-Lingual Coreference Resolution

Although supervised machine learning methods have consistently driven model advancements, these rely on large annotated coreference corpora, and thus cannot be applied to most of the world's languages. The task of cross-lingual coreference resolution has recently gathered attention, benefiting from several shared tasks.

Cross-lingual knowledge transfer generally relies on either a common multi-lingual semantic space [49,96,97] or on parallel corpora, which enables projection of information from one language to another [98,99]. The use of multilingual word embeddings [49] as a multi-lingual semantic space enables the use of direct transfer learning between languages [100]. In a direct transfer approach, the system is trained on a source language, where annotated data is available, and then the learned model is used to initialize a new model that will work on a target language. The model can be used to form predictions on the target data (without being explicitly trained using labeled data on the target language) after updating the embedding layer for the target language using multilingual word embeddings. Consequently, a model trained on one language can be used for any other language that shares its semantic space [28,29]. Cruz et al. [28] explored the direct transfer approach to leverage a Spanish corpus for coreference resolution in the Portuguese language. They reported competitive results compared to an in-language model, which supports further exploring transfer learning techniques to address less-resourced languages using the proposed approach. Similarly, Kundu et al. [29] showed that a model trained on English and tested on Chinese and Spanish achieved results competitive with those obtained by models trained directly on Chinese and Spanish, respectively. As an extrinsic evaluation task for the proposed cross-lingual coreference model, they showed that the English model helps achieve entity-linking accuracy on Chinese and Spanish test sets higher than the top TAC 2015 trilingual entity discovery and linking (EDL) task [15] system without using any annotated data from Chinese or Spanish. As the mapping of different languages into a common semantic space is still error-prone [49], truly language-agnostic systems still do not exist; however, improving multilingual embeddings is bound to improve cross-lingual system performance.

The TAC-KBP shared tasks [15–17] tackled cross-lingual EDL between Chinese, Spanish, and English, with the latest task also featuring 10 additional low-resource languages. Recent advances have dramatically increased cross-language portability in this task, with current techniques being able to perform fine-grained name tagging and linking on hundreds of languages [101].

Several works have used projection-based techniques [102,103] for improving performance on less-resourced languages [98,99,104,105]. In projection approaches, learning instances originally in the source language are translated, using machine translation tools or parallel data, to the target language, and the corresponding labels are projected to the new learning instances in the target language. Fine-grained word alignment techniques [106] can be employed to produce high quality translations and to better preserve the annotation's token-level boundaries. Then, the model is trained on the projected data directly in the target language. Rahman and Ng [104] used machine translation to apply projection-based approaches for multilingual coreference resolution, presenting promising results on two target languages: Spanish and Italian. Martins [98] applied projection on an English-Portuguese-Spanish parallel corpus, reporting strong results compared with other cross-lingual methods, such as bilingual word embeddings, bitext direct projection, or vanilla posterior regularization. For the same goal, Novák et al. [99] reported improved cross-lingual results on the CORBON 2017 shared task [36], building German and Russian coreference resolvers based solely on parallel English data.

The approach proposed by Howard and Ruder [107] for universal language model fine-tuning for text classification could further improve current performance on less-resourced languages, promising to achieve NLP transfer learning gains similar to those achieved by computer-vision models, and showing promising results on other challenging NLP tasks (sentiment analysis, question classification, and topic classification). Recently proposed multilingual contextualized word embeddings, including mBERT [44] (the multilingual counterpart of BERT), are a promising line of work in this direction, without any reported systematic evaluation on cross-lingual settings at the moment of this writing, to the best of our knowledge.

7. Common Challenges

Due to its old roots, coreference resolution has several challenges well-documented in the literature. This section describes and explores these common challenges and predicaments, most of which are not exclusive to coreference resolution systems.

7.1. Biases

The biases reflected in classifiers are a growing concern in the machine learning community, and coreference resolution is no exception. These biases mainly come from two sources, training data and auxiliary resources, and affect rule-based systems and neural systems alike [108].

Rule-based systems rely on corpus-based gender/number statistics mined from external resources [109], providing counts for the frequency with which a noun phrase is observed in a male, female, neutral, and plural context, thus reflecting the biases inherent to the given corpus. Rudinger et al. [108] experimented with artificially balancing these frequencies, significantly decreasing the biases in the system at the cost of a small performance decline.

Word embeddings, which are the basis of current state-of-the-art neural systems [6,26,43,44], are also known to be severely gender biased [110,111]. Bolukbasi et al. [110] reported clear gender-based stereotypes on pre-trained embeddings in their paper entitled “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”, and experimented with debiasing techniques. Caliskan et al. [48] extensively demonstrated that embeddings trained on human language result in human-like semantic biases. The authors used the Implicit Association Test (IAT) [112] to test for several known human biases on GloVeword embeddings [113]. Caliskan et al. first replicated standard implicit associations between flowers-pleasant and insects-unpleasant, then showing, among other trends, the association between male-career and female-family. Recently, Zhao et al. [111] presented GN-GloVe, a gender-neutral version of GloVe [113] word embeddings.

Along these lines, Zhao et al. [114] developed a dataset (WinoBias) for evaluating the bias in coreference resolution systems. The OntoNotes 5.0 dataset, used by virtually all state-of-the-art systems [5], is known to have male mention heads in over 80% of mentions headed by a gendered pronoun. As such, state-of-the-art systems link gendered pronouns to pro-stereotypical entities with significantly higher accuracy than anti-stereotypical entities, with an average 21.1 F1 score difference, as measured by Zhao et al. [114]. Webster et al. [115] presented the GAPdataset, a manually-annotated corpus of ambiguous pronoun-name pairs derived from Wikipedia snippets, serving as a benchmark for gender bias in coreference resolution. The authors found systematic performance disparities between genders in state-of-the-art systems. Namely, the system constructed by Lee et al. [13] achieves a bias-factor of 0.89, meaning performance on female entities is 89% that of male entities. The current state-of-the-art system by Joshi et al. [6] achieves a bias-factor of 0.95, reducing the gender bias to promising levels.

Additionally, datasets are susceptible to annotator biases, as it is not always clear for humans which expressions corefer. The ACE 2003 dataset [31] reports an inter-annotator agreement of 52% on its English relation detection and characterization (RDC) portion, whereas the OntoNotes dataset, a more recent corpus, publicizes an inter-annotator agreement of over 90% [27].

As shown by the recency of the references in this section, researchers are growing increasingly aware of their models' and dataset biases, and this seems a productive and impactful topic for future research.

7.2. Imbalanced Datasets

As a mention is only coreferent with a small subset of other mentions from its possibly much higher number of candidates, the generated data used to train supervised models typically feature heavily imbalanced datasets. This is a prevailing problem in coreference resolution, and a well studied problem in machine learning in general [69,116,117].

A common approach to address this problem is to artificially change the data distribution [118], either by oversampling or undersampling (or a combination of the two). Oversampling consists of synthetically creating instances of the minority class by means of replicating existent data instances or by using some artificial combination of their features as new training instances. Undersampling consists of removing training samples of the majority class. Fonseca et al. [69] explored random undersampling in coreference resolution, with encouraging results. Rocha and Lopes Cardoso [119] proposed heuristic-based strategies to undersample the originally unbalanced dataset of mention-pair learning instances. These training set creation strategies explore well-known properties of coreference resolution to generate more balanced distribution of labels while providing suitable learning instances for the mention-pair models. For instance, the most confident antecedent neighbors (MCAN) strategy, instead of generating a negative example for each mention that occurs between a coreferent mention-pair, generates up to k negative examples with the antecedents occurring closer to the coreferent mention-pair. Using these heuristic-based strategies, the overall performance of the system is improved compared with the random undersampling approach.

Unfortunately, these techniques have their weaknesses, as oversampling tends to lead to overfitting, and undersampling may deprive the model of useful training instances [28,119].

7.3. World Knowledge

As suggested by many researchers [1,13], truly solving coreference resolution would require comprehensive world knowledge, along with common-sense reasoning. This has typically been addressed by using extensive external APIs, such as WordNet [39] and BabelNet [120]. Over the years, there have been several publications on this topic, mainly regarding joint tackling coreference resolution and entity linking, using the linked knowledge-base to improve coreference performance [4,10]. Sapena et al. [1] reported small but consistent improvements using WordNet to judge word similarity, resulting in improved recall of coreference links but with decreased precision, which the authors postulated to be due to the typical noise produced by WordNet. By exploring semantic-based features (based on WordNet and word embeddings similarity metrics), Rocha and Lopes Cardoso [119] improved the performance of the system compared to some baseline methods for the Portuguese language, mainly because these features allowed the system to solve some coreferences between noun phrases (e.g., semantically similar words, synonyms, hyperonyms, meronyms) that were not captured by lexical-, syntactical-, morphological-, or structural-based features.

In the last few years, the increasing popularity of deep learning led to most coreference resolution systems using word embeddings to encode world knowledge and semantic relations between words. As previously pointed out, Peters et al. [43] presented contextualized word embeddings, allowing for a word's vector representation to also depend on its context, resulting in significant performance improvements for several NLP tasks (including coreference resolution). Along these lines, BERT contextual embeddings were recently introduced, improving the performance boundary of several

NLP tasks, including coreference resolution [6]. Recent studies showed that these contextualized word embeddings, such as BERT, pre-trained on huge amounts of unlabeled data, are able to recall factual and common sense knowledge at a level remarkably competitive with non-neural and supervised alternatives [121], suggesting that these approaches provide a good alternative to embed world knowledge in the system. An additional advantage is that learning these embeddings does not require manually annotated data, acting as general multitask learners [47]. Aiming to ease transfer learning between different languages, Grave et al. [49] released multilingual word vectors for 157 languages, opening up new possibilities for language-agnostic models.

7.4. Mention Detection

As discussed in Section 2.1, mention detection is often considered part of the coreference resolution task. Recent shared tasks evaluate system performance on system-detected mentions [27]. As such, an improvement in mention detection would be immediately reflected in the coreference resolution performance. This step generally aims for a high recall to the detriment of precision, as wrongly identified mentions are often classified as non-coreferent (singletons), but missed mentions may provoke error propagation and will more severely affect coreference metrics through the mention identification effect [50].

As current state-of-the-art coreference resolution is performed by end-to-end systems, jointly tackling mention detection and linking shows an increasing consideration of the mention detection phase. Lee et al. [13] reported a one point improvement on the CoNLL metric with joint mention detection and mention scoring over their baseline system, and a 17.5-point CoNLL gain when using oracle mentions (gold standard mentions instead of automatically identified ones), suggesting room for improvement along these lines of research.

8. Conclusions

In this work, we explored the state-of-the-art advances in the field of coreference resolution, and their interwoven relationship with common NLP tasks. Despite the continuous improvements over the years, a variety of challenges need to be addressed in order to create a truly domain-agnostic coreference resolution system.

Although it is evidently mandatory to incorporate world knowledge to advance in this field, current systems focus on entity-level similarities and reasoning, lacking common sense inference, as seen in the “Winograd Schema” examples. Global entity-level information is also a requirement for solving most pronominal mention links, giving cluster-ranking or higher-order models more credit for future research. The use of dense vector representations (embeddings) to encode mentions and their real-world relationships has yielded remarkable results, managing to surpass the performance of previous systems based on fine-tuned and hand-engineered features. This is consistent with similar radical improvements seen in all areas of machine learning due to the advent of deep learning. Additionally, in the case of natural language processing, the introductions of BERT and the transformer architecture have expanded the capabilities in most fields of language understanding, and coreference resolution is no exception.

End-to-end differentiable systems that jointly tackle more than one task have also considerably improved coreference resolution performance, whether jointly tackling this task with mention detection, quotation attribution, or entity-linking. These tasks provide the model extra information that is useful for the coreference task, mainly when addressing mention detection, as it was traditionally part of a pipeline system for coreference resolution. State-of-the-art works regularly report performance for a single model and a model ensemble, showing that ensembles consistently provide an additional performance boost.

The reliance on supervised machine learning methods presents a challenge for low-resource languages. Although several transfer learning approaches have demonstrated promising results, further work on transfer learning from higher-resourced languages is needed to increase coreference resolution performance on languages that lack a large annotated dataset for this task. Additionally, better and context-dependent multilingual semantic spaces are likely to aid in this task. Work on unsupervised and semi-supervised methods is also necessary when large multilingual semantic spaces are not available for a given language.

As a final remark, an end-to-end differentiable system able to capture mention boundaries, coreference links, and their entity-level transitivity, but also reason over the entities' roles in a text, is a promising way forward for coreference resolution and deeper language understanding.

Author Contributions: Conceptualization and methodology, G.R. and Henrique L.C.; investigation and writing—original draft preparation, A.F.C.; writing—review and editing, G.R. and H.L.C.; supervision and funding acquisition, H.L.C. All authors have read and agreed to the published version of the manuscript.

Funding: André Ferreira Cruz is supported by the Calouste Gulbenkian Foundation, under grant number 214508. Gil Rocha is supported by a PhD studentship (with reference SFRH/BD/140125/2018) from Fundação para a Ciência e a Tecnologia (FCT). This research is partially supported by project DARGMINTS (POCI/01/0145/FEDER/031460), funded by FCT.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Sapena, E.; Padró, L.; Turmo, J. A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution. *Comput. Linguist.* **2013**, *39*, 847–884. [CrossRef]
2. Levesque, H.J.; Davis, E.; Morgenstern, L. The Winograd schema challenge. In Proceedings of the AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning, Palo Alto, CA, USA, 21–23 March 2011; Volume 46, p. 47.
3. Rahman, A.; Ng, V. Coreference Resolution with World Knowledge. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 814–824.
4. Hajishirzi, H.; Zilles, L.; Weld, D.S.; Zettlemoyer, L. Joint Coreference Resolution and Named-Entity Linking with Multi-Pass Sieves. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 289–299.
5. Ng, V. Machine Learning for Entity Coreference Resolution: A Retrospective Look at Two Decades of Research. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4877–4884.
6. Joshi, M.; Levy, O.; Zettlemoyer, L.; Weld, D. BERT for Coreference Resolution: Baselines and Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 5802–5807. [CrossRef]
7. Poesio, M.; Stuckardt, R.; Versley, Y. *Anaphora Resolution: Algorithms, Resources, and Applications*; Springer: Berlin/Heidelberg, Germany, 2016.
8. Sukthanker, R.; Poria, S.; Cambria, E.; Thirunavukarasu, R. Anaphora and Coreference Resolution: A Review. *arXiv* **2018**, arXiv:1805.11824.
9. Almeida, M.S.; Almeida, M.B.; Martins, A.F. A Joint Model for Quotation Attribution and Coreference Resolution. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, 26–30 April 2014; pp. 39–48.
10. Durrett, G.; Klein, D. A Joint Model for Entity Analysis: Coreference, Typing, and Linking. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 477–490. [CrossRef]
11. Nadeau, D.; Sekine, S. A survey of named entity recognition and classification. *Lingvist. Investig.* **2007**, *30*, 3–26.
12. Daumé III, H.; Marcu, D. A Large-Scale Exploration of Effective Global Features for a Joint Entity Detection and Tracking Model. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, 6–8 October 2005; pp. 97–104.

13. Lee, K.; He, L.; Lewis, M.; Zettlemoyer, L. End-to-end Neural Coreference Resolution. In Proceedings of the 2017 Conference on Empirical Methods in NLP, Copenhagen, Denmark, 7–11 September 2017; pp. 188–197.
14. Ji, H.; Nothman, J.; Hachey, B. Overview of TAC-KBP2014 Entity Discovery and Linking Tasks. In Proceedings of the Text Analysis Conference (TAC2014), Gaithersburg, MD, USA, 17–18 November 2014; pp. 1333–1339.
15. Ji, H.; Nothman, J.; Hachey, B.; Florian, R. Overview of TAC-KBP2015 Tri-lingual Entity Discovery and Linking. In Proceedings of the Eighth Text Analysis Conference (TAC2015), Gaithersburg, MD, USA, 16–17 November 2015.
16. Ji, H.; Nothman, J.; Dang, H.T.; Hub, S.I. Overview of TAC-KBP2016 Tri-lingual EDL and Its Impact on End-to-End Cold-Start KBP. In Proceedings of the TAC, Gaithersburg, MD, USA, 14–15 November 2016.
17. Ji, H.; Pan, X.; Zhang, B.; Nothman, J.; Mayfield, J.; McNamee, P.; Costello, C.; Hub, S.I. Overview of TAC-KBP2017 13 Languages Entity Discovery and Linking. In Proceedings of the Tenth Text Analysis Conference (TAC2017), Gaithersburg, MD, USA, 13–14 November 2017.
18. Voutilainen, A. Part-of-Speech Tagging. In *The Oxford Handbook of Computational Linguistics*; Oxford University Press: Oxford, UK, 2003; pp. 219–232.
19. Toutanova, K.; Klein, D.; Manning, C.D.; Singer, Y. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, AB, Canada, 27 May–1 June 2003; pp. 173–180.
20. Akbik, A.; Blythe, D.; Vollgraf, R. Contextual String Embeddings for Sequence Labeling. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1638–1649.
21. Bohnet, B.; McDonald, R.; Simões, G.; Andor, D.; Pitler, E.; Maynez, J. Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2642–2652.
22. Marcus, M.P.; Santorini, B.; Marcinkiewicz, M.A. Building a Large Annotated Corpus of English: The Penn Treebank. *Comput. Linguist.* **1993**, *19*, 313–330.
23. Yang, X.; Su, J.; Tan, C.L. Kernel-Based Pronoun Resolution with Structured Syntactic Knowledge. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17–21 July 2006; pp. 41–48.
24. Bengtson, E.; Roth, D. Understanding the Value of Features for Coreference Resolution. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; pp. 294–303.
25. Durrett, G.; Klein, D. Easy Victories and Uphill Battles in Coreference Resolution. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1971–1982.
26. Lee, K.; He, L.; Zettlemoyer, L. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In Proceedings of the 2018 Conference of the NAACL: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; Volume 2, pp. 687–692.
27. Pradhan, S.; Moschitti, A.; Xue, N.; Uryupina, O.; Zhang, Y. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In Proceedings of the Joint Conference on EMNLP and CoNLL-Shared Task, Jeju Island, Korea, 12–14 July 2012; pp. 1–40.
28. Ferreira Cruz, A.; Rocha, G.; Lopes Cardoso, H. Exploring Spanish Corpora for Portuguese Coreference Resolution. In Proceedings of the Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), Valencia, Spain, 15–18 October 2018.
29. Kundu, G.; Sil, A.; Florian, R.; Hamza, W. Neural Cross-Lingual Coreference Resolution And Its Application To Entity Linking. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 2, pp. 395–400.
30. Hirschman, L.; Chinchor, N. Appendix F: MUC-7 Coreference Task Definition (version 3.0). In Proceedings of the Seventh Message Understanding Conference (MUC-7), Fairfax, VA, USA, 29 April–1 May 1998.

31. Doddington, G.; Mitchell, A.; Przybocki, M.; Ramshaw, L.; Strassel, S.; Weischedel, R. The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004), Lisbon, Portugal, 26–28 May 2004.
32. Recasens, M.; Màrquez, L.; Sapena, E.; Martí, M.A.; Taulé, M.; Hoste, V.; Poesio, M.; Versley, Y. Semeval-2010 task 1: Coreference resolution in multiple languages. In Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, 15–16 July 2010; pp. 1–8.
33. Chaimongkol, P.; Aizawa, A.; Tateisi, Y. Corpus for Coreference Resolution on Scientific Papers. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland, 26–31 May 2014; pp. 3187–3190.
34. Schäfer, U.; Spurk, C.; Steffen, J. A Fully Coreference-annotated Corpus of Scholarly Papers from the ACL Anthology. In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012): Posters, Mumbai, India, 8–15 December 2012; pp. 1059–1070.
35. Cohen, K.B.; Lanfranchi, A.; Choi, M.J.y.; Bada, M.; Baumgartner, W.A.; Panteleyeva, N.; Verspoor, K.; Palmer, M.; Hunter, L.E. Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. *BMC Bioinform.* **2017**, *18*, 372. [CrossRef]
36. Grishina, Y. CORBON 2017 Shared Task: Projection-Based Coreference Resolution. In Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017), Valencia, Spain, 4 April 2017; pp. 51–55.
37. Nedoluzhko, A.; Novák, M.; Ogrodniczuk, M. PAWS: A Multi-lingual Parallel Treebank with Anaphoric Relations. In Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference, New Orleans, USA, 6 June 2018; pp. 68–76.
38. Sundheim, B.M. Overview of Results of the MUC-6 Evaluation. In Proceedings of the 6th Conference on Message Understanding (MUC-6), Association for Computational Linguistics, Columbia, MD, USA, 6–8 November 1995; pp. 13–31. [CrossRef]
39. Miller, G.A. WordNet: A lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
40. Vossen, P. EuroWordNet: A Multilingual Database of Autonomous and Language-specific Wordnets Connected via an Inter-Lingual-Index. *Int. J. Lexicogr.* **2004**, *17*, 161–173. [CrossRef]
41. Navigli, R.; Ponzetto, S.P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.* **2012**, *193*, 217–250. [CrossRef]
42. Clark, K.; Manning, C.D. Deep Reinforcement Learning for Mention-Ranking Coreference Models. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 2256–2262.
43. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. 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), New Orleans, LA, USA, 1–6 June 2018; Volume 1, pp. 2227–2237.
44. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
45. Rubenstein, H.; Goodenough, J.B. Contextual correlates of synonymy. *Commun. ACM* **1965**, *8*, 627–633. [CrossRef]
46. Socher, R.; Chen, D.; Manning, C.D.; Ng, A. Reasoning With Neural Tensor Networks for Knowledge Base Completion. *NIPS Proc.* **2013**, *1*, 926–934.
47. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. Available online: https://openai.com/blog/better-language-models/ (accessed on 29 December 2019).
48. Caliskan, A.; Bryson, J.J.; Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **2017**, *356*, 183–186. [CrossRef] [PubMed]
49. Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; Mikolov, T. Learning Word Vectors for 157 Languages. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.

50. Moosavi, N.S.; Strube, M. Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 632–642.
51. Vilain, M.; Burger, J.; Aberdeen, J.; Connolly, D.; Hirschman, L. A Model-theoretic Coreference Scoring Scheme. In Proceedings of the 6th Conference on Message Understanding (MUC-6), Columbia, MD, USA, 6–8 November 1995; pp. 45–52. [CrossRef]
52. Luo, X. On Coreference Resolution Performance Metrics. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, BC, Canada, 6–8 October 2005; pp. 25–32. [CrossRef]
53. Bagga, A.; Baldwin, B. Algorithms for Scoring Coreference Chains. In Proceedings of the 1st International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference, Granada, Spain, 28–30 May 1998; Volume 1, pp. 563–566.
54. Stoyanov, V.; Gilbert, N.; Cardie, C.; Riloff, E. Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, Suntec, Singapore, 2–7 August 2009; pp. 656–664.
55. Rahman, A.; Ng, V. Supervised Models for Coreference Resolution. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, Singapore, 6–7 August 2009; pp. 968–977.
56. Cai, J.; Strube, M. Evaluation Metrics For End-to-End Coreference Resolution Systems. In Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Tokyo, Japan, 24–25 September 2010; pp. 28–36.
57. Recasens, M.; Hovy, E. BLANC: Implementing the Rand index for coreference evaluation. *Nat. Lang. Eng.* **2011**, *17*, 485–510. [CrossRef]
58. Kobdani, H.; Schütze, H. SUCRE: A Modular System for Coreference Resolution. In Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, 15–16 July 2010; pp. 92–95.
59. Rand, W.M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* **1971**, *66*, 846–850. [CrossRef]
60. Luo, X.; Pradhan, S.; Recasens, M.; Hovy, E. An extension of BLANC to system mentions. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Baltimore, MD, USA, 22–27 June 2014; pp. 24–29.
61. Kantor, B.; Globerson, A. Coreference Resolution with Entity Equalization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 673–677. [CrossRef]
62. Ma, X.; Hovy, E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 1064–1074.
63. Luo, X.; Ittycheriah, A.; Jing, H.; Kambhatla, N.; Roukos, S. 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), Barcelona, Spain, 21–26 July 2004.
64. Klenner, M.; Ailloud, É. Enhancing Coreference Clustering. In Proceedings of the Second Workshop on Anaphora Resolution, Bergen, Norway, 29–31 August 2008; pp. 31–40.
65. Cai, J.; Strube, M. End-to-End Coreference Resolution via Hypergraph Partitioning. In Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, Beijing, China, 23–27 August 2010; pp. 143–151.
66. Clark, K.; Manning, C.D. Entity-Centric Coreference Resolution with Model Stacking. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, 26–31 July 2015; pp. 1405–1415.
67. Ng, V.; Cardie, C. Improving Machine Learning Approaches to Coreference Resolution. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 104–111.

68. Ng, V. Unsupervised Models for Coreference Resolution. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; pp. 640–649.
69. Fonseca, E.B.; Vieira, R.; Vanin, A. Dealing with Imbalanced Datasets for Coreference Resolution. In Proceedings of the Twenty-Eighth International Flairs Conference, Hollywood, FL, USA, 18–20 May 2015.
70. Finkel, J.R.; Manning, C.D. Enforcing Transitivity in Coreference Resolution. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Columbus, OH, USA, 15–20 June 2008; pp. 45–48.
71. Denis, P.; Baldridge, J. Specialized models and ranking for coreference resolution. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; pp. 660–669.
72. Wiseman, S.; Rush, A.M.; Shieber, S.; Weston, J. 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), Beijing, China, 26–31 July 2015; pp. 1416–1426.
73. Wiseman, S.; Rush, A.M.; Shieber, S.M. Learning Global Features for Coreference Resolution. In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, June 12–17 2016; pp. 994–1004.
74. Clark, K.; Manning, C.D. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016;, pp. 643–653.
75. Gasperin, C. Semi-supervised anaphora resolution in biomedical texts. In Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language Processing and Biology, New York, NY, USA, 8 June 2006; pp. 96–103.
76. Raghavan, P.; Fosler-Lussier, E.; Lai, A.M. Exploring Semi-Supervised Coreference Resolution of Medical Concepts using Semantic and Temporal Features. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montréal, QC, Canada, 3–8 June 2012; pp. 731–741.
77. Blum, A.; Mitchell, T. Combining Labeled and Unlabeled Data with Co-Training. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, Madison, WI, USA, 24–26 July 1998; pp. 92–100.
78. Haghighi, A.; Klein, D. Unsupervised Coreference Resolution in a Nonparametric Bayesian Model. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 23–30 June 2007; pp. 848–855.
79. Poon, H.; Domingos, P. Joint Unsupervised Coreference Resolution with Markov Logic. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; pp. 650–659.
80. Teh, Y.W.; Jordan, M.I.; Beal, M.J.; Blei, D.M. Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. *NIPS Proc.* **2015**, *1*, 1385–1392.
81. Lee, L.S.; Chien, L.F.; Lin, L.J.; Huang, J.; Chen, K.J. An Efficient Natural Language Processing System Specially Designed for the Chinese Language. *Comput. Linguist.* **1991**, *17*, 347–374.
82. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc.* **1977**, *39*, 1–22.
83. Ma, X.; Liu, Z.; Hovy, E. Unsupervised Ranking Model for Entity Coreference Resolution. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1012–1018.
84. Fei, H.; Li, X.; Li, D.; Li, P. End-to-end Deep Reinforcement Learning Based Coreference Resolution. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 660–665. [CrossRef]
85. Zhang, R.; Nogueira dos Santos, C.; Yasunaga, M.; Xiang, B.; Radev, D. Neural Coreference Resolution with Deep Biaffine Attention by Joint Mention Detection and Mention Clustering. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia, 15–20 July 2018; pp. 102–107.

86. Martschat, S.; Strube, M. Latent Structures for Coreference Resolution. *Trans. Assoc. Comput. Linguist.* **2015**, *3*, 405–418. [CrossRef]
87. Björkelund, A.; Kuhn, J. 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), Baltimore, MD, USA, 22–27 June 2014; Volume 1, pp. 47–57.
88. Dietterich, T.G. Ensemble Methods in Machine Learning. In *International Workshop on Multiple Classifier Systems*; Springer: Berlin/ Heidelberg, Germany, 2000; pp. 1–15.
89. Singh, S.; Riedel, S.; Martin, B.; Zheng, J.; McCallum, A. Joint Inference of Entities, Relations, and Coreference. in Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, San Francisco, CA, USA, 27–28 October 2013; pp. 1–6.
90. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
91. Graves, A. Generating Sequences with Recurrent Neural Networks. *arXiv* **2013**, arXiv:1308.0850.
92. Dozat, T.; Manning, C.D. Deep Biaffine Attention for Neural Dependency Parsing. In Proceedings of the 5th International Conference on Learning Representations (ICLR 2017), Toulon, France, 24–26 April 2017.
93. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *NIPS Proc.* **2019**, *1*, 5754–5764.
94. Conneau, A.; Lample, G. Cross-lingual Language Model Pretraining. *NIPS Proc.* **2019**, *1*, 7057–7067.
95. Kiela, D.; Bhooshan, S.; Firooz, H.; Testuggine, D. Supervised Multimodal Bitransformers for Classifying Images and Text. *arXiv* **2019**, arXiv:1909.02950.
96. Camacho-Collados, J.; Pilehvar, M.T.; Navigli, R. A Unified Multilingual Semantic Representation of Concepts. 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), Beijing, China, 26–31 July 2015; pp. 741–751.
97. Cao, Y.; Huang, L.; Ji, H.; Chen, X.; Li, J. Bridge Text and Knowledge by Learning Multi-Prototype Entity Mention Embedding. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1623–1633.
98. Martins, A.F. Transferring Coreference Resolvers with Posterior Regularization. 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), Beijing, China, 26–31 July 2015; pp. 1427–1437.
99. Novák, M.; Nedoluzhko, A.; Žabokrtskŷ, Z. Projection-based Coreference Resolution Using Deep Syntax. In Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017), Valencia, Spain, 4 April 2017; pp. 56–64.
100. McDonald, R.; Petrov, S.; Hall, K. Multi-source Transfer of Delexicalized Dependency Parsers. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011), Edinburgh, UK, 27–31 July 2011; pp. 62–72.
101. Pan, X.; Zhang, B.; May, J.; Nothman, J.; Knight, K.; Ji, H. Cross-lingual Name Tagging and Linking for 282 Languages. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1946–1958.
102. Yarowsky, D.; Ngai, G.; Wicentowski, R. Inducing Multilingual Text Analysis Tools via Robust Projection Across Aligned Corpora. In Proceedings of the First International Conference on Human Language Technology Research, San Diego, CA, USA, 18–21 March 2001; pp. 1–8. [CrossRef]
103. Hwa, R.; Resnik, P.; Weinberg, A.; Cabezas, C.; Kolak, O. Bootstrapping Parsers via Syntactic Projection Across Parallel Texts. *Nat. Lang. Eng.* **2005**, *11*, 311–325. [CrossRef]
104. Rahman, A.; Ng, V. Translation-Based Projection for Multilingual Coreference Resolution. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montréal, QC, Canada, 3–8 June 2012; pp. 720–730.
105. Grishina, Y.; Stede, M. Knowledge-lean projection of coreference chains across languages. In Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, Beijing, China, 30 July 2015; pp. 14–22.
106. Dyer, C.; Chahuneau, V.; Smith, N.A. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 644–648.

107. Howard, J.; Ruder, S. Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 328–339.
108. Rudinger, R.; Naradowsky, J.; Leonard, B.; Van Durme, B. Gender Bias in Coreference Resolution. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), New Orleans, LA, USA, 1–6 June 2018; pp. 8–14.
109. Bergsma, S.; Lin, D. Bootstrapping Path-Based Pronoun Resolution. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17–21 July 2006; pp. 33–40.
110. Bolukbasi, T.; Chang, K.W.; Zou, J.Y.; Saligrama, V.; Kalai, A.T. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *NIPS Proc.* **2016**, *1*, 4349–4357.
111. Zhao, J.; Zhou, Y.; Li, Z.; Wang, W.; Chang, K.W. Learning Gender-Neutral Word Embeddings. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 4847–4853.
112. Greenwald, A.G.; McGhee, D.E.; Schwartz, J.L. Measuring individual differences in implicit cognition: The implicit association test. *J. Personal. Soc. Psychol.* **1998**, *74*, 1464. [CrossRef]
113. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
114. Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; Chang, K.W. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), New Orleans, LA, USA, 1–6 June 2018; pp. 15–20.
115. Webster, K.; Recasens, M.; Axelrod, V.; Baldridge, J. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 605–617. [CrossRef]
116. Mazurowski, M.A.; Habas, P.A.; Zurada, J.M.; Lo, J.Y.; Baker, J.A.; Tourassi, G.D. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Netw.* **2008**, *21*, 427–436. [CrossRef]
117. Chawla, N.V. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 875–886.
118. More, A. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv* **2016**, arXiv:1608.06048.
119. Rocha, G.; Lopes Cardoso, H. Towards a Mention-Pair Model for Coreference Resolution in Portuguese. In *EPIA Conference on Artificial Intelligence*; Springer: Cham, Switzerland, 2017; Volume 10423, pp. 855–867.
120. Moro, A.; Cecconi, F.; Navigli, R. Multilingual Word Sense Disambiguation and Entity Linking for Everybody. In Proceedings of the International Semantic Web Conference (Posters & Demos), Riva del Garda, Italy, 19–23 October 2014; pp. 25–28.
121. Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 2463–2473. [CrossRef]