Artificial Intelligence Review (2023) 56:14439–14481
https://doi.org/10.1007/s10462-023-10506-3

A brief survey on recent advances in coreference resolution

Ruicheng Liu[1] · Rui Mao[1] · Anh Tuan Luu[1] · Erik Cambria[1]

Accepted: 12 May 2023 / Published online: 26 May 2023

Abstract
The task of resolving repeated objects in natural languages is known as coreference resolution, and it is an important part of modern natural language processing. It is classified into two categories depending on the resolved objects, namely entity coreference resolution and event coreference resolution. Predicting coreference connections and identifying mentions/triggers are the major challenges in coreference resolution, because these implicit relationships are particularly difficult in natural language understanding in downstream tasks. Coreference resolution techniques have experienced considerable advances in recent years, encouraging us to review this task in the following aspects: current employed evaluation metrics, datasets, and methods. We investigate 10 widely used metrics, 18 datasets and 4 main technical trends in this survey. We believe that this work is a comprehensive roadmap for understanding the past and the future of coreference resolution.

Keywords Coreference resolution · Natural language processing · Artificial intelligence · Deep learning

1 Introduction

A collection of statements that have a logical structure and a consistent meaning when taken together is referred to as a discourse. To achieve coherence within the discourse, it is necessary to have a firm grasp of the argumentation structure and information flow. Coreference resolution is among these parsing attempts and anaphora resolution is a subset of it. The resolution of anaphoras is the process of determining the antecedents of referring

✉ Erik Cambria
cambria@ntu.edu.sg

Ruicheng Liu
ruicheng001@e.ntu.edu.sg

Rui Mao
rui.mao@ntu.edu.sg

Anh Tuan Luu
anhtuan.luu@ntu.edu.sg

[1] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

phrases. Coreference resolution (CR) is a wider definition that refers to the process of resolving any spans in a context that point to the same physical object or event.

Coreference resolution is important for downstream natural language processing activities such as entity linking (Kundu et al. 2018), named entity recognition (Dai et al. 2019), question answering (Bhattacharjee et al. 2020), sentiment analysis (Krishna et al. 2017; Mao and Li 2021) and chatbots (Zhu et al. 2018). It also has strong connections in referring expression generations (Li et al. 2018; Chen et al. 2018). The purpose of this study is to offer a quick review of recent advances in addressing the coreference resolution issues. The models reviewed in this survey are categorized into four classes: feature-based approaches, multilayer perceptron/ recurrent neural network approaches, knowledge-based approaches, and transformer-based approaches. Feature-based approaches concentrate on leveraging lexical, grammatical, and semantic information. Multilayer perceptron/ recurrent neural network approaches are end-to-end approaches that use neural network models to understand the contextual information of mentions but they do not employ external knowledge explicitly. Knowledge-based neural approaches are those that explicitly employ external knowledge, and are generally constructed on top of multilayer perceptron/ recurrent neural approaches. Transformer-based approaches have sparked attention in the NLP field in recent years (Devlin et al. 2019; Liu et al. 2019; Yang et al. 2019; Lan et al. 2019). Specifically, models built on top of BERT and SpanBERT have shown exceptional performance in terms of coreference resolution. Pre-trained language models can be viewed as neural networks that have incorporated commonsense knowledge and contextual information implicitly through complex embeddings.

In previous surveys, Mitkov (1999) proposed one of the earliest works focusing on the explanation of anaphora resolution and different algorithms. Ng (2010) presented how Machine Learning started aiding coreference resolution in the first 15 years of this field. Lu and Ng (2018) focused on event coreference resolution field and summarized the research works from 1997 to 2017 including supervised, semi-supervised, and unsupervised approaches. However, it did not cover entity coreference resolution and only touched briefly on neural network-based approaches. Sukthanker et al. (2020) summarized more recent development in coreference and anaphora resolution and showed some deep learning approaches. However, there are new datasets and models introduced in the coreference resolution field in recent years that are not covered by Sukthanker et al. (2020). Especially after the introduction of Transformers (Vaswani et al. 2017), Transformer-based large-scale pre-trained language models have brought natural language processing research into a new era. This motivates us to deliver a survey of coreference resolution, covering up-to-date evaluation metrics, datasets, and technical trends.

To bridge this gap, this survey aims at providing readers with a rough timeline of how coreference resolution has evolved from feature-based and classical machine learning-based approaches to deep learning-based approaches. For deep learning-based approaches, they are further divided into neural-based contextual, neural-based knowledge, and transformer-based approaches. We will provide insight into each technical trend, relevant datasets, and evaluation metrics. Last but not least, the summarization in this paper goes into further details than previous survey works, for example, we presented a summary of tools for coreference annotation (Sect. 5), a summary of application-oriented datasets (Table 1), methodology summaries (Tables 2 and 3), feature summaries (Table 4), and external knowledge summaries (Table 5).

In the remaining sections of this survey, we start by defining the two common coreference resolution types and tasks in Sect. 2, followed by the most frequently used measure in Sect. 3 and datasets in Sect. 4. Next, we investigate four main types of coreference resolution models from Sects. 6 to 9 and summarize learning methods, features and external knowledge used by those models in Sect. 10. Furthermore, we analyze the challenges of current coreference resolution methods and provide constructive recommendations for future works in Sect. 11. Finally, we conclude the survey in Sect. 12.

2 Coreference resolution types and tasks

2.1 Entity coreference resolution

Entity mentions are spans of words that can be used to represent real world entities. Entity Coreference Resolution (ECR) is a task that groups entity mentions into sets of mentions that refer to the same real-world entities. An antecedent of a mention is the entity to whom the anaphoric words refer to. When there are two or more mentions referring to the same antecedent, we say these mentions corefer. A singleton, on the other hand, is a mention that is only mentioned once in a document.

Here is an example regarding entity coreference resolution:

> *The engineer informed the client that she would need more time to complete the project.*

In the sentence above, there are 5 mentions that could represent real world entities: *The engineer*, *the client*, *she*, *time* and *the project*. Among these 5 mentions, *The engineer* and *she* are referring to the same real world entity, therefore *The engineer* and *she* are coreferential. As the pronoun *she* is pointing to a mention that appears before it, *she* is an anaphoric mention and *The engineer* is its antecedent. The other 3 mentions are singletons.

2.2 Event coreference resolution

Unlike entity mentions, event mentions consist of multiple textual spans, including an event anchor and multiple arguments. An event anchor could be a verb, gerund or noun while the arguments refer to the subject and object if applicable. Event coreference resolution groups together event mentions that refer to the same event. The event mentions can be contained inside a single document (denoted as within document) or spread over several documents (denoted as cross document). It is critical for information aggregation and can assist a variety of downstream natural language processing applications, such as contradiction detection (de Marneffe et al. 2008), text summarization (Ferracane et al. 2016), and reading comprehension (Khashabi et al. 2018; Welbl et al. 2018).

Consider the following example that was also shown in Lu and Ng (2021c):

> *Yesterday the Delhi Police {slapped }$_{ev1}$ a protester while she was {demonstrating }$_{ev2}$ outside a hospital. At almost the same time, a woman in her 60 s was {beaten up }$_{ev3}$ by policemen in another {protest }$_{ev4}$ in the northern Indian state of Uttar Pradesh. As of now, the Delhi Police has suspended the cop who {assaulted }$_{ev5}$ the woman protester.*

In the snippet above, there are five event mentions ($ev1 - ev5$), which are "*slapped*", "*demonstrating*", "*beaten up*", "*protest*", and "*assaulted*", respectively. Each event mention (trigger word) could have its own arguments (subjects and objects). e.g. in the event "the Delhi Police slapped a protester", "the Delhi Police" and "a protester" are two arguments of the trigger word "*slapped*". While *ev*1, *ev*3, and *ev*5 are of subtype ATTACK, only *ev*1 and *ev*5 are coreferent, as *ev*3 took place during a different protest. In addition, *ev*2 and *ev*4 are not coreferent because they refer to different PROTEST events.

2.3 Coreference resolution tasks

Both entity coreference resolution and event coreference resolution tasks are implemented as a pipeline consisting of mention[1] detection and mention linking tasks.

Mention detection is a critical component of entity coreference resolution. It was observed that mention detection might restrict the performance of the coreference resolver (Poesio et al. 2016). Conventionally, mention detection is evaluated separately as a stand alone task in order to properly compare different models (Lu and Ng 2020).

Mention linking is the task of clustering the detected mentions to match the gold standard. Some coreference resolution model performed both mention detection and mention linking jointly (Lee et al. 2018; Joshi et al. 2020), whereas some models separated the two tasks and only focused on linking given mentions (Khosla and Rose 2020; Caciularu et al. 2021).

3 Metrics

In many commonly used datasets, such as GAP (Webster et al. 2018), DPR (Rahman and Ng 2012), WSC (Levesque et al. 2012), Winogender (Rudinger et al. 2018) and PDP (Davis et al. 2017), coreference resolution problems can be treated as word-level binary classification problems. These datasets are prepared in a gold-two-mention style, containing paired sentences, the first of which has two or more mentions, and the second of which contains an ambiguous pronoun. A model should link the ambiguous pronoun to the correct mention. In this case, precision, recall, and F1 score (Sect. 3.1) are widely used for measuring both mention detection and mention linking evaluation.

However, coreference resolution problem can go beyond gold-two-mention problems because it usually includes clustering mentions into multiple coreference clusters and each cluster could contain multiple mentions (on the contrary, gold-two-mention problems only have one coreference link, which is the pronoun and its linked mention from the two candidate mentions). When mention linking is evaluated in general coreference resolution tasks, specialized metrics such as MUC (Sect. 3.2), B-Cubed (Sect. 3.3), CEAF (Sect. 3.4), BLANC (Sect. 3.5), and LEA (Sect. 3.6) are employed. In this section, we also present evaluation metrics with special purposes (Sect. 3.7), e.g., metrics for measuring the gender bias of coreference resolution systems. Finally, we present the typical combinations of individual measures (Sect. 3.8) in evaluating coreference resolution.

3.1 F1 score

It is common to use precision, recall and F1 score to evaluate the mention detection (Yu et al. 2020; Peng et al. 2015) and binary selection-based mention linking tasks (Kocijan et al. 2019; Attree 2019). They are the most widely used performance measures in a binary classification task. Precision is defined as number of true positive predictions divided by number of all positive predictions:

[1] either as entity mentions or event mentions.

$$precision = \frac{|true\ positives|}{|true\ positives| + |false\ positives|}, \tag{1}$$

where $|\cdot|$ denotes the number of items. Number of true positive predictions divided by number of actual positive items is referred to as recall:

$$recall = \frac{|true\ positives|}{|true\ positives| + |false\ negatives|}. \tag{2}$$

F1 score relates to the harmonic mean of precision and recall, which is given by

$$F_1 = \frac{2}{recall^{-1} + precision^{-1}}. \tag{3}$$

Sometimes, accuracy is also reported for measuring performance (Rudinger et al. 2018; Zhao et al. 2018; Kocijan et al. 2019). It refers to the ratio of the number of properly anticipated items (the sum of true positive and true negative predictions) to the total number of items

$$accuracy = \frac{|true\ positives| + |true\ negatives|}{|total\ items|}. \tag{4}$$

3.2 MUC

MUC is the earliest coreference evaluation measure that was introduced by Vilain et al. (1995). MUC is a measure that is based on links. Links are coreferential relations between mentions. If two mentions corefer, there is a link between them. We define K as the key set which is a set of mentions that is clustered in the correct way. Each cluster within K is denoted as $k \in K$. All the mentions within the same cluster k are co-referential according to the hard truth (gold standard). R denotes the response set which is the set of mentions clustered by an evaluated model. r denotes one of the cluster within response set R. Then, the MUC Precision value is computed as below:

$$MUC\,Precision(K,R) = \sum_{r \in R} \frac{|r| - |partition(r,K)|}{|r| - 1}, \tag{5}$$

where $|r|$ denotes the total number of mentions within cluster r, $|partition(r,K)|$ denotes the number of segments induced in the response cluster r in relation to the key clusters in K. It is formed by intersecting r with each key cluster $k \in K$ that overlaps with r. For example, if the mentions within a response cluster r belongs to 5 different key clusters $k \in K$, then $|partition(r,K)| = 5$, which means this response cluster r can be partitioned by K into 5 segments. We refer readers to Vilain et al. (1995) for more details regarding this calculation.

Similar to MUC Precision, the MUC Recall value is computed as below:

$$MUC\,Recall(K,R) = \sum_{k \in K} \frac{|k| - |partition(k,R)|}{|k| - 1}, \tag{6}$$

where $|k|$ denotes the total number of mentions within cluster k, $|partition(k, R)|$ denotes the count of segments of key cluster k relative to response set R. Each partition is formed by intersecting k and those response set $r \in R$ that overlaps with k (Vilain et al. 1995).

3.3 B^3(B-Cubed)

The B^3 score was introduced by Bagga and Baldwin (1998). B^3 is a mention-based measure, i.e., the overall recall or precision is calculated by using the recall or precision of individual mentions. For each mention m_i, B^3 recall examines the proportion of overlapped mentions in both the key cluster (K_i) containing mention m_i and the response cluster (R_i) containing mention m_i above the number of mentions in the key cluster (K_i) containing mention m_i. B^3 recall for mention m_i is computed as follows:

$$Recall_i = \frac{|K_i \cap R_i|}{|K_i|} \tag{7}$$

Similarly, B^3 precision for mention i is computed by changing the key clusters to response clusters in the denominator:

$$Precision_i = \frac{|K_i \cap R_i|}{|R_i|} \tag{8}$$

The final B^3 precision and recall are the weighted sum of individual entity scores:

$$\begin{aligned} Precision &= \sum_{i=1}^{N} w_i * Precision_i \\ Recall &= \sum_{i=1}^{N} w_i * Recall_i. \end{aligned} \tag{9}$$

Usually the weights (w_i) are assigned with $1/N$, where N represents the total number of mentions to be considered.

3.4 CEAF

The Constrained Entity Alignment F-measure (CEAF) proposed by Luo (2005) is used for entity or mention-based similarity detection. CEAF first creates a one-to-one mapping between response clusters and key clusters based on similarity. It then calculates accuracy and recall using this mapping. Luo (2005) provided four distinct forms of the similarity assessments:

$$\phi_1(K, R) = \begin{cases} 1 & \text{if } R = K \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

$$\phi_2(K, R) = \begin{cases} 1, & \text{if } R \cap K \neq \emptyset \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

$$\phi_3(K, R) = |R \cap K| \tag{12}$$

$$\phi_4(K,R) = 2 \cdot \frac{|R \cap K|}{|R + K|} \tag{13}$$

The key entities are represented by K, while the response entities are represented by R. $\phi_1(K,R)$ asserts that two entities are the same only if they share all the mentions, whereas $\phi_2(K,R)$ asserts that two entities are the same as long as they share at least a mention. $\phi_3(\mathrm{K},\mathrm{R})$ is the number of shared mentions between key clusters and response clusters, whereas $\phi_4(\mathrm{K},\mathrm{R})$ represents the number of shared mentions relative to the size of key clusters and response clusters.

CEAF comes in two flavors: mention-based and entity-based. The function $m(k)$ maps each key cluster k to a response cluster r using the Kuhn-Munkres algorithm (Kuhn 1955). The precision and recall of mention-based $CEAF_m$ are specified as follows:

$$CEAF_m\,Precision(K,R) = \frac{\sum_{k_i \in K} \phi\big(k_i, m(k_i)\big)}{\sum_{r_i \in R^*} |r_i|}, \tag{14}$$

$$CEAF_m\,Recall(K,R) = \frac{\sum_{k_i \in K} \phi\big(k_i, m(k_i)\big)}{\sum_{k_i \in K} |k_i|}, \tag{15}$$

where ϕ could be any function from ϕ_1 (Eq. 10) to ϕ_4 (Eq. 13), whereas ϕ_3 (Eq. 12) and ϕ_4 are most commonly used (Luo 2005). $|r_i|$ represents the total number of mentions within cluster r_i. $|k_i|$ represents the total number of mentions within cluster k_i. R^* represents the subset of response entities that can be mapped to K.

The precision and recall of entity-based $CEAF_e$ are computed as:

$$CEAF_e\,Precision(K,R) = \frac{\sum_{k_i \in K} \phi\big(k_i, m(k_i)\big)}{N_r}, \tag{16}$$

$$CEAF_e\,Recall(K,R) = \frac{\sum_{k_i \in K} \phi\big(k_i, m(k_i)\big)}{N_k}, \tag{17}$$

where N_r represents the total number of response entities and N_k represents the total number of key entities.

3.5 BLANC

BLANC (Recasens and Hovy 2011) is a link-based measure that is based on rand indices. It looks at coreference links and non-coreference links separately. Recall and precision of coreference links are computed as:

$$R_c = \frac{|C_k \cap C_r|}{|C_k|}, \quad P_c = \frac{|C_k \cap C_r|}{|C_r|}, \tag{18}$$

where C_k represents the coreference links in the key clusters and C_r represents the coreference links in the response clusters.

Recall and precision of non-coreference links are computed as:

$$R_n = \frac{|N_k \cap N_r|}{|N_k|}, \quad P_n = \frac{|N_k \cap N_r|}{|N_r|},$$

where, N_k represents the non-coreference links in the key clusters and N_r represents the non-coreference links in the response clusters.

Final BLANC recall and precision are the average scores by coreference and non-coreference links

$$Recall = \frac{R_c + R_n}{2},$$
$$Precision = \frac{P_c + P_n}{2}.$$

3.6 LEA

In LEA (Moosavi and Strube 2016), the proportion of successfully resolved connections between mentions is used to compute recall. The amount of mentions for each entity is weighted in the results, such that successfully resolving an entity with more mentions contributes more to the total score. Precision is calculated by inverting the key and response cluster roles.

$$\text{Recall} = \frac{\sum_{k_i \in K} \left[|k_i| \times \sum_{r_j \in R} \frac{\text{link}(k_i \cap r_j)}{\text{link}(k_i)} \right]}{\sum_{k_i \in K} |k_i|}$$
$$\text{Precision} = \frac{\sum_{r_i \in R} \left[|r_i| * \sum_{k_j \in K} \frac{\text{link}(r_i \cap k_j)}{\text{link}(r_i)} \right]}{\sum_{r_z \in R} |r_z|}$$

where for any cluster S, link(S) denotes the total number of edges of a complete graph with each node representing a mention from the same cluster. $\text{link}(S) = |S| \times (|S| - 1)/2$

3.7 Special aspects of analysis on coreference systems

There are other miscellaneous metrics which focus on certain specialized aspects, such as gender bias in coreference resolution. Zhao et al. (2018) created a new benchmark dataset called WinoBias, who measured the difference between pro-stereotyped and anti-stereotyped scenarios (e.g. in a woman dominated profession, linking a female pronoun with the job name is considered as 'pro-stereotypical', and linking a male pronoun with that job is considered as 'anti-stereotypical'). A robust coreference resolver should be able to handle both scenarios well. Besides that, the performance difference under the two scenarios should not be significant.

Similarly, Emami et al. (2019) proposed a corpus that switches candidate antecedents with different gender and number cues in order to mislead coreference resolvers. An outstanding system which relies on knowledge and contextual information should not be misled by such kind of lexical changes. The *consistency* score is thus defined as the proportion of correct predictions with the modified sentences in the corpus.

Varkel and Globerson (2020) and Wu et al. (2020) also used a bias factor. The bias factor is defined as F_1^f / F_1^m. It is the ration of F1 on feminine examples (*f*) and F1 on masculine examples (m).

If neither of the mentions in the gold-two-mention task is taken into account, the task is formalized as a three-class classification problem. Abzaliev (2019) used logarithmic loss to assess the model performance. The loss is given by

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log\left(p_{i,j}\right),$$

where N denotes the number of test set samples. M denotes number of classes ($M = 3$). $y_{i,j}$ is an indicator function. It takes value 1, if observation i belongs to class j. Otherwise, it takes value 0. $p_{i,j}$ denotes the predicted probability that i belongs to class j.

3.8 Combination of measures

There is no single metric that is universally suitable in the coreference resolution domain, due to the complexity of the task. It is common to incorporate several evaluation metrics together in the CR research. B-cubed, MUC and CEAF are the three most commonly used evaluation metrics in both entity and event coreference resolution tasks. Each of them can be formalized in the form of precision, recall and F1 score measures, respectively, where F1 scores are most commonly chosen as the overall measure. For example, Joshi et al. (2020) compared SpanBERT-based models, BERT-based models and end-to-end models with regard to F1 scores of MUC, B-cubed and $CEAF_{\phi_4}$ respectively. The final performance is measured by the averaged F1 score. This average score is also called the CoNLL score, as it was used by the CoNLL 2012 shared task (Pradhan et al. 2012; Huang et al. 2019). Lu and Ng (2021a) further took BLANC into account when computing the average of F1 scores over event coreference tasks. The average of the four scores is also known as AVG-F. Additionally, for a more comprehensive overview of the performance of a coreference resolution system, detection precision, detection recall and detection F-measure could be also employed in coreference resolution performance evaluation (Zhang et al. 2018).

3.9 Summary of evaluation metrics

For gold-two-mention style coreference resolution tasks, F1 and accuracy are the standard evaluation metrics as they are quite intuitive. For higher order coreference resolutions, concerns have been expressed by researchers regarding the assessment measures that are used. Despite recent model advances, the CoNLL score remains the key evaluation measure employed by state-of-the-art models in recent years, which uses the F1 average of *MUC*, B^3 and *CEAF*. However, these three measures all have their pitfalls. *MUC* is considered to have the weakest ability to differentiate good and bad coreference resolution results (Recasens and Hovy 2011). It also prefers coreference result that is over-merged (Luo 2005). B^3 may lead to counter-intuitive results under some edge cases (Luo 2005). It cannot handle repeated mentions very well (Luo and Pradhan 2016). $CEAF_e$ treat mention clusters equally irrespective of their sizes (Stoyanov et al. 2009). Additional measurements have been created in recent years to overcome the limitations of these three traditional metrics. Agarwal et al. (2019) established new measures for evaluating name entity coreference (NEC) after determining that existing metrics did not meet the criteria of the NEC task. Moosavi and Strube (2016) introduced the LEA measure to account for the importance of entities with greater mentions. In addition to traditional measurements, researchers are advised to consider using these new metrics as well.

4 Datasets

4.1 CoNLL 2012 data

CoNLL 2012 shared task (Pradhan et al. 2012) proposed three coreference resolution datasets in English, Chinese and Arabic. The datasets were built upon OntoNotes v5.0 (Hovy et al. 2006), containing texts from different sources, e.g., broadcast, magazine, newswire, weblogs and newsgroups. The texts may involve multiple speakers, e.g., in broadcast and telephone conversations. Alternatively, the texts are monologues. The English dataset of CoNLL-2012 shared task includes 2802 training documents, 343 development documents and 348 testing documents. The Chinese version dataset consists of 1810 training documents, 252 validation documents and 218 test documents. The Arabic version dataset consists of 359 training documents, 44 validation documents and 44 test documents. To the best of our knowledge, the current state-of-the-art model is Wang et al. (2021) with CoNLL score of 87.5%.

4.2 GAP

The GAP dataset (Webster et al. 2018) was sourced from Wikipedia snippets. Each snippet is annotated with one gender-ambiguous pronoun, two names, and two flags. A model has to decide which name the gender-ambiguous pronoun refers to. The model is then evaluated, based on the coreference connections between the two names and the pronoun. F1 scores on masculine, feminine and overall examples are commonly used metrics on the GAP dataset. And the ratio F_1^f/F_1^m (F1 score for feminine examples over F1 score for masculine examples) is also calculated to evaluate the gender bias. GAP dataset contains 8908 pairs of ambiguous pronouns and candidate mentions. The training, validation and testing snippets have 4000, 908, and 4000 samples, respectively. The current state-of-the-art on GAP test dataset is the ProBERT (Attree 2019) with F1 score of 92.5% and 0.97 gender bias.

4.3 KBP 2017 event coreference dataset

The TAC KBP Event Track dataset (Mitamura et al. 2017) is used to resolve event coreference in the TAC KB 2017 shared task. The goal of the TAC KBP Event track is to extract information about events such that the information could be used as inputs into a knowledge base. Event Nugget (EN) Detection, Coreference, and Sequencing tasks, as well as Event Argument and Linking (EAL) tasks in the shared task are evaluated at the document level. Except for event sequencing, all other event tasks are in three languages, namely English, Chinese, and Spanish. The KBP 2017 shared task provided a standard measure, AVG-F. AVG-F is the average of four widely used metrics: MUC, B^3, $CEAF_e$ and BLANC (see Sect. 3). KBP 2017 dataset contains 167 documents. The state-of-the-art performance is introduced in the work of Yu et al. (2020) with an AVG-F of 57.12%.

4.4 ACE2005

The Linguistic Data Consortium (LDC) created the ACE (Automatic Content Extraction) 2005 Multilingual Training Corpus, which comprises about 1,800 files of mixed genre texts with annotations in entities, relations, and events in English, Arabic, and Chinese. ACE2005 is the whole collection of the training data with various languages in the 2005 ACE technology evaluation. The genres cover the texts from newswire, broadcast news, broadcast conversation, weblog, discussion forums, and conversational telephone voice. LDC annotated the data with assistance from the ACE Program and additional assistance from LDC. ACE2005-English contains 599 files, ACE2005-Chinese contains 633 files and ACE2005-Arabic contains 403 files. The state-of-the-art performance on ACE2005-English is introduced by Lai et al. (2021) with a CoNLL score of 87.90% and AVG-F score of 88.30%.

4.5 LitBank

LitBank is a new dataset of literary text coreferences (Bamman et al. 2020). The collection contains 100 lary texts with an average length of about 2100 words. Singletons are recognized and evaluated. The original evaluation was based on 10-fold cross validation with 80%, 10%, and 10% data splits for training, validation and testing. It restricts the mentions to six entity categories (location, organizations, people, vehicles, geo-political entities, facilities) with the bulk of mentions (83.1%) pointing to entities belonging to the people category. Khosla and Rose (2020) introduced the state-of-the-art performance with CoNLL score of 80.26% on this dataset.

4.6 WSC

The Winograd Schema Challenge (WSC) (Levesque 2011) is a hard pronoun resolution challenge based on Winograd's (Winograd 1972) examples.

A Winograd Schema example reflects the situation where a single word modification in a sentence changes the referent of the pronoun, making the resolution difficult. The goal is to determine which entity the pronoun or possessive adjective refers to in a context. The context includes two entities. The text contains a "special word". The statement remains technically valid when the "alternative word" is used for substitution, whereas the referent of the pronoun changes. Consider the example below:

> *William could only climb beginner walls while Jason climbed advanced ones because he was very [weak/strong].*

In this sentence, *weak* is a special word while *strong* is an alternative word. When *weak* is used in this sentence, pronoun *he* will refer to *William*. If *strong* is used to substitute *weak*, then the pronoun *he* will refer to *Jason* instead.

The Winograd Schema Challenge consists of challenging cases that need commonsense to answer. These cases could not be solved simply using statistical analysis of co-occurrences and associations. The SuperGLUE (Wang et al. 2019) version of WSC dataset contains 554 training examples, 104 validation examples and 146 test examples. The state-of-the-art model is claimed to be ERNIE 3.0 (Sun et al. 2021) with accuracy of 97.3%.

4.7 DPR

The Definite Pronoun Resolution (DPR) corpus (Rahman and Ng 2012) is a modified version of Winograd Schema Challenge-style issues. These sentence pairs span a wide range of themes, from real occurrences to cinematic events to entirely fictitious circumstances, primarily representing pop culture as experienced by American children born in the early 1990 s. DPR includes cases that do not need commonsense reasoning, as well as situations where the "special word" is a phrase. DPR contains 1322 training examples and 564 test examples. Totally, there are 1886 example sentences. The state-of-the-art model on DPR dataset is the BERT_WIKICREM_ALL (Kocijan et al. 2019) with an accuracy of 84.8%.

4.8 PDP

The Pronoun Disambiguation Problem (PDP) (Davis et al. 2017) is a modest set of 60 questions that served as the first round of the 2016 Winograd Schema Challenge. Unlike WSC, the cases do not involve a "special word" in PDP. However, they still need commonsense thinking to understand the texts. The samples were hand-picked from literature. The state-of-the-art model for PDP is BERT_WIKICREM_ALL (Kocijan et al. 2019) with an accuracy of 86.7%.

4.9 Winogender

Winogender (Rudinger et al. 2018) is a dataset for testing the gender biases in coreference resolution, using the WSC format. Each sentence has an occupational noun and a referring pronoun. The pronoun could be represented as "he", "she" or "they", respectively. The occupational nouns are usually gender-oriented. E.g., women are likely to be employed as secretaries. Given "the secretary asked the visitor to sign in so that he could update the guest log" (Rudinger et al. 2018), a coreference resolution classifier may fail in connecting "he" to "secretary", if the classifier is gender-biased. This dataset means to examine how altering the gender of the pronoun impacts the accuracy of a model. Winogender contains 720 sentences in total. The state-of-the-art model on this dataset is BERT_WIKICREM_DPR (Kocijan et al. 2019) with accuracy of 82.1%.

4.10 WinoBias

WinoBias (Zhao et al. 2018) is also a WSC-inspired dataset that measures gender biases in coreference resolution algorithms. Similar to Winogender, WinoBias contains examples of occupations with a high gender imbalance. It contains 3160 Winograd Schemas examples, equally divided into training and test sets. The test set examples are divided into two types, where Type 1 examples are prototypical WSC phrases. Coreference judgments must utilize world knowledge based on the given conditions. Such instances are difficult to understand because they lack syntactic clues. Type 2 examples utilize syntactic knowledge and a pronoun comprehension. Since both semantic and syntactic clues aid in disambiguation, resolvers are likely perform better in Type 2 instances. The gender of the pronominal reference is immaterial for the co-reference judgment in both types. To pass the test, systems must be able to produce valid linkage predictions in both pro- and anti-stereotypical circumstances. The stereotyped jobs were chosen using data from the US Department of Labor. The best performance is introduced by BERT_DPR (Kocijan et al. 2019) on Type 1 subset (with accuracy of 78.0%–78.2%) and BERT_WIKICREM_ALL (Kocijan et al. 2019) on Type 2 subset (with accuracy of 98.7%–99.0%).

4.11 KnowRef

Emami et al. (2019) introduced KnowRef, a coreference resolution corpus that particularly tests the capacity of a system to reason about a scenario stated in the context.

KnowRef is a human-labeled corpus with 8,724 Winograd-like text samples, the resolution of which necessitates considerable commonsense and domain knowledge. Each instance consists of a brief text with a target pronoun that must be appropriately resolved to one of two potential antecedents. The KnowRef dataset was created by collecting text samples from a vast collection of documents, including 2018 English Wikipedia, OpenSubtitles, and Reddit comments. KnowRef contains 7455 training sentences and 1269 testing sentences. The state-of-the-art model on KnowRef is BERT(KnowRef) (Emami et al. 2019).

4.12 WikiCoref

WikiCoref (Ghaddar and Langlais 2016) includes annotated Wikipedia documents. Documents were carefully chosen to span a variety of stylistic articles. Each mention is annotated with entity type and coreference properties, as well as the Freebase subject to which it belongs. The annotation scheme of WikiCoref is the extension of the OntoNotes scheme. WikiCoref consists of 30 documents with an average document size of 2000 tokens. Khosla and Rose (2020) held the best performance with a CoNLL score of 71.35%.

4.13 ECB+

Extension to Event Coreference Bank (ECB+) (Cybulska and Vossen 2014) consists of within- and cross-document coreference annotations for entities and events. The identification of groupings of related texts that describe the same foundational event is a key stage in the construction of the ECB+ corpus, enabling for the annotation of coreferential event references across documents. Different topics from Google News archives were chosen in order to contain intentionally selected keywords. ECB+ contains 976 documents in total which are divided into 574 documents for training, 196 documents for validation and 206 documents for testing. The current state-of-the-art model for ECB+ is Cross Document Language Model (CDLM) (Caciularu et al. 2021) with a CoNLL score of 85.6%.

4.14 Richer event description (RED)

Richer Event Description (RED) corpus (O'Gorman et al. 2016) annotates entities, events, and times, as well as their coreference connections and the temporal, causal, and subevent linkages between the events. It contains 8731 events, 1127 temporal expressions, and 10320 entities in 95 documents (totaling 54287 tokens), sampled from both news data and casual discussion forum interactions. It includes 2390 identity chains, 1863 bridging relations, and 4969 event-event relations that include temporal, causal, and subevent relationships, as well as 8731 DocTimeRel temporal annotations that connect these events to the document time.

4.15 Georgetown university multilayer corpus (GUM)

Georgetown University Multilayer corpus (GUM) (Zeldes 2017) was collected in the context of classroom teaching. It includes rich annotated texts of twelve genres from various sources including Wikinews, Wikivoyage, Wikihow, Reddit. Main annotations in this corpus include multiple Part-of-Speech (POS) tags, document structure in TEI XML (paragraphs, headings, figures, etc.), constituent and dependency syntax, entity and coreference annotation, discourse dependencies. It includes 168 documents with 150824 tokens.

4.16 WEC

The Wikipedia Event Coreference (Eirew et al. 2021) is a data set for a cross-document event coreference task. Data annotation is boosted by leveraging available information in Wikipedia while the coreferences are not restricted by predefined topics. The information is gathered by grouping together the anchor texts of (internal) Wikipedia links pointing to the same Wikipedia concept. This is typically justified because all of these links are about the same real-world subject. As a result, the WEC dataset is made up of mentions, each of which contains the mention span corresponding to the link anchor text, the surrounding context, and the mention cluster ID. Since Wikipedia was not divided into predefined topics, mentions can have coreference links across the entire corpus. WEC training set consists of 40529 event mentions in 7042 clusters. The validation set consists of 1250 event mentions in 233 clusters. The test set consists of 1893 event mentions in 322 clusters. The state-of-the-art model is introduced in Eirew et al. (2021) with a CoNLL score of 62.3%.

4.17 EmailCoref

EmailCoref (Dakle et al. 2020) includes 46 email threads and 245 email messages. This is the first dataset to address the problem of entity resolution in email threads. It has set two rules for choosing email threads: The thread must have at least three email messages, with at least half of the email messages including text content. EmailCoref contains 36 training email threads and 10 testing email threads. Khosla and Rose (2020) introduced the best performance with a CoNLL score of 76.17%.

4.18 BUG

Levy et al. (2021) presented BUG, a large scale gender bias dataset that has similar challenging style as Winogender (Rudinger et al. 2018) and WinoBias (Zhao et al. 2018). BUG was semi-automatically collected with help of SPIKE (Shlain et al. 2020) from three different sources: Wikipedia, PubMed and Covid19 research papers. BUG has 108K sentences and the state-of-the-art model is the SpanBERT fine tuned on anti-stereotypical part of BUG with an accuracy of 64.1%.

4.19 Annotation formatting

There are two mainstream annotation formats for coreference resolution datasets: CoNLL format (Hovy et al. 2006) and Winograd format (Levesque et al. 2012). CoNLL formatted datasets mainly include datasets of CoNLL 2012, WikiCoref, GUM and EmailCoref. In CoNLL format, every word is represented in one line and sentences are separated with blank lines. CoNLL format can have multiple columns with each column representing one type of annotation, in the coreference resolution annotation column, selected mentions are labeled with cluster id, if two mentions are coreferent with each other then they will have the same cluster id. Gold-two-mention type of datasets mainly includes GAP, WSC, DPR, WikiCREM, PDP, Winogender, WinoBias and KnowRef. These dataset are labelled in the way that is identical or similar to the Winograd format (Levesque et al. 2012), i.e. they will provide the position of the pronoun/mask and the two candidate mentions in the sentence and they will also label the correct choice of the two candidate mentions. Some datasets also got their own special types of annotation tags, e.g. in ECB+ (Cybulska and Vossen 2014), *INTRA_DOC_COREF* and *CROSS_DOC_COREF* tags are employed to capture within-document coreference and cross document coreference chains respectively.

5 Annotation tools

There are several tools that can be used for CR annotation.

BRAT (Stenetorp et al. 2012) is a web-based tool that supports manual annotation for a variety of NLP tasks such as chunking, dependency syntax and coreference resolution. BRAT is based on a client–server architecture. The backend is implemented in Python. It also integrates a semantic class disambiguation component in order to reduce ambiguity and help retain a correct class (Stenetorp et al. 2011).

MMAX2 (Kopeć 2014) is a desktop-based coreference annotation tool, written in Java. The main features of MMAX2 include: the semantic head word of a mention; the attribute selection of a cluster; merging two mentions into one with one click; a plugin that allows users to see the differences between the two versions of annotations and merge them into one.

CorefAnnotator (Reiter 2018) is an open source desktop application, distributed under license Apache 2.0. Equivalence sets are used to represent coreference chains. Each entity is represented by a color, and can be named, if desired. In the text view, all mentions of the same object are underlined with the same color; numerous annotations on the same span result in multiple underlines on different levels. It allows importing and exporting in a few file formats, including Excel for easy analysis.

SACR (Oberle 2018) is a web-based application that works with Firefox or Google Chrome. It is available online, while it can also be downloaded and used offline. Mentions are identified by clicking on their first and last words, whereas coreference relationships are established by dragging one mention and dropping it on another one.

COREFI (Bornstein et al. 2020) is web-component tool that can be embedded into any website. It is designed for crowdsourcing setting. The whole process includes three parts: onboarding, annotation and review. The onboarding component includes a dialogue-based tutorial and a guided annotation task that allows users to be familiar with the functions of the tool. In the annotation process, annotators are prompted with candidate mentions one at a time. The annotators can decide whether to adjust the boundaries of the mentions, and which cluster to assign the mention to. All the operations can be done via keyboard operations. In the reviewing process, the reviewer is presented with the candidate clusters, annotated by annotators and the reviewer can modify the spans and clusters.

In summary, most of the annotation tools are able to perform annotation for both mention identification and mention linking, the annotation processes are slightly different among different tools. In pair-based tools such as BRAT (Stenetorp et al. 2012) and MMAX2 (Kopeć 2014), annotators will first determine the mention' boundaries and link mentions one pair at a time. In cluster-based tools such as CorefAnnotator (Reiter 2018), SACR (Oberle 2018) and COREFI (Bornstein et al. 2020), annotators will first determine the mentions' boundaries and then drop mentions into respective clusters. Some tools such as BRAT (Stenetorp et al. 2012) are meant for general purpose annotations with different NLP tasks rather than coreference resolution only.

6 Feature-based approaches

Feature-based models include linguistic information such as part-of-speech (POS) and named-entity recognition (NER) tags, as well as surface-level semantic information, such as opinion words. The models themselves are also not based on deep learning, but rather on standard Machine Learning approaches (e.g. decision trees, memory-based learning and conditional random fields). These models represent the early stages of coreference resolution research. They are typically fairly intuitive, whereas they are incapable of capturing deep level contextual information and comparatively lack generalization capacity.

The objects and attributes of coreference resolution issues, proposed by Ding and Liu (2010), are the challenge of detecting whether references of objects and attributes correspond to the same entities. To tackle the problem, they employed a supervised learning approach. The major contribution of the article is the creation and testing of two unique opinion-related characteristics for learning. The first feature was based on non-comparative sentence sentiment analysis, comparative sentence sentiment analysis, and the idea of sentiment consistency. The second feature took into account which objects and attributes were modified by which opinion words. Opinion words, such as *good*, *best*, *bad*, and *poor*, are often used to convey positive or negative feelings. Their model workflow included preprocessing, feature vector construction, classifier construction, and testing. The model first preprocessed the corpus by running a POS tagger and a Noun Phrase finder. They then generated the object-noun phrase (O-NP) set, which includes potential objects, attributes, and other noun phrases. Then, for each pair in the O-NP set, they created a feature vector. Since their study focused on products and attributes, they left out personal pronouns, gender agreement features, and appositive features. Training data were created in the classifier construction step with each pair containing at least one object or attribute. To fit the training data, a decision tree was built. Several novel features in the opinion mining context were proposed in this work, including sentiment consistency, object/attribute, and opinion word associations.

Atkinson et al. (2015) combined features-based coreferencing and memory-based learning which improves opinion retrieval in social media. The working model was built on top of three main tasks: massage retrieval, message preprocessing and reference analysis, and opinion retrieval. Message retrieval collected and stored the hierarchies of various tweets in a local database. Tokenization, POS tagging and named-entity identification were used to extract essential underlying linguistic information from the collected tweets. In the message referencing analysis and opinion retrieval stage, a memory-based learning approach (MBL) was utilized. A Machine Learning approach that searches for the training data item that is most similar to the test data item and make predictions based on the similarity is referred to as an MBL. As major generalization approaches, memory-based learning systems employed nearest-neighbor search, space decomposition techniques, and clustering. This feature-based referencing classification model was tested using formal and informal text corpora. The results showed that the accuracy for extracting referential links on the formal texts improved more compared with the informal texts, due to the linguistic features of informal messages.

A joint model of three essential activities for the entity analysis stack was provided by Durrett and Klein (2014): coreference resolution, the identification of entities and the entity linking. The joint model took unary, binary and ternary factors into account when solving these three problems. Unary factors were features employed when solving each task in isolation. Binary and ternary factors were introduced to capture cross-task interactions. For example, the restriction of coreferential references having the same semantic kind. Based on Durrett and Klein (2014)'s original argument for jointly modeling, namely that the three tasks have possible impacts on each other, they showed that making use of the interactions between the modules resulted in higher performance overall. As a result, any pipelined system would inevitably underperform a combined model.

The mention-ranking technique to coreference was used in Durrett and Klein (2014). Their feature set focused on the surface features of mentions, such as starting and ending word, mention length, and the syntactic role of each mention. Coreference features incorporated multiple features between mention pairs as well as aspects of the mention pair itself, such as distance between mentions and whether their heads matched. Anaphora features explored each of these qualities in turn.

Raghunathan et al. (2010a) applied a multi-level sieve structure that applied one sieve at each level, the sieve with the higher precision will always come before the lower precision sieve. This design aims to avoid the phenomenon of lower precision feature prevailing over the higher precision feature.

In summary, common features that are widely used in feature-based approaches are opinion words (Ding and Liu 2010), POS tags (Ding and Liu 2010; Atkinson et al. 2015), text chunking (Ding and Liu 2010), NER tags (Atkinson et al. 2015; Raghunathan et al. 2010a), semantic information (Durrett and Klein 2014), syntactic roles (Durrett and Klein 2014), word positions and head words (Durrett and Klein 2014; Raghunathan et al. 2010a). Feature-based approaches mainly represent the early stage of coreference resolution studies. Their performance has been exceeded by the latest deep learning-based approaches. Since feature-based approaches are not the focus of coreference resolution research community in the past decade, we only list a few approaches, published after 2010. If readers are interested in very early studies, one can refer to the survey of Ng (2010).

7 Multilayer perceptron/recurrent neural network approaches

Neural-based models are trained using neural networks to understand the contextual information of natural language input and abstract features in a high dimensional vector space. Generally these models do not include external knowledge besides the training dataset itself. For disambiguation, this section mainly refers to the neural network models in the pre-BERT era.

7.1 Entity-centric CR with model stacking

Clark and Manning (2015) showed how mention pair model scores can be combined to generate powerful entity-level properties between mention clusters. Using these properties, an entity-centric coreference system was trained to develop an appropriate policy for progressively building up coreference chains. The scores obtained by mention pair models were used as features in training an entity-centric system. The majority of task-specific learning happened inside mention pair models, which were trained via a simple supervised method. The entity-centric agent then learned an efficient technique for progressively building up coreference clusters, utilizing prior decisions to influence future ones, guided by the pairwise scores.

The entity-centric system constructed coreference chains via agglomerative clustering (Clark and Manning 2015): each mention began in its own cluster, and pairs of clusters were merged at each stage. By using an imitation learning method based on DAgger (Ross et al. 2011), an agent was trained to decide if it was beneficial to combine a specific pair of clusters (Ross et al. 2011). The model employed a method of attributing actual costs to actions based on coreference assessment metrics, as well as a perception of the gravity of an error. Furthermore, rather than evaluating all pairings of clusters as candidate merges, the pairwise model scores were utilized to narrow the search space, first by giving an ordering over which merges were evaluated, and then by rejecting merges that were unlikely to be correct. This significantly decreased the time required to operate the agent, making learning computationally viable.

7.2 Learning global features for CR with RNN

Wiseman et al. (2016) claimed that global context is required for future advances in coreference resolution. However, it is difficult to design informative cluster-level characteristics, which limits their utilities. As a result, Wiseman et al. (2016) proposed to employ a Recurrent Neural Network (RNN) to train representations of mention clusters sequentially. The model did not have any explicit clustering characteristics; instead, from the individual mentions in each cluster, it developed a global representation. These representations were implemented into a style coreference system based on mention-ranking. The whole model, including the RNN and the mention-ranking sub-system, was trained from end to end on the coreference dataset.

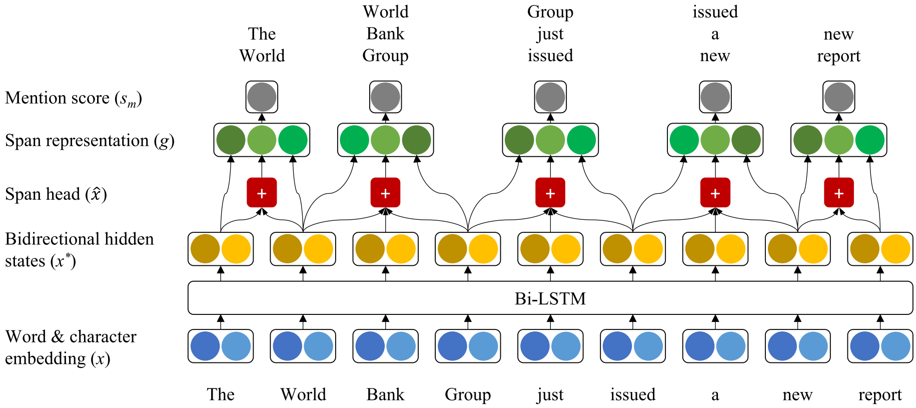

Fig. 1 The span representation generated by using bidirectional LSTMs and mention scores. The figure is adapted from the work of Lee et al. (2017)

7.3 End-to-end neural CR

Lee et al. (2017) introduced the first end-to-end CR model and showed that it outperforms every prior efforts without using any syntactic parser or an explicit mention detector. The basic idea was to take all spans in a text as possible mentions and learn the probability distribution over the antecedents for the spans. By integrating contextual-based boundary representations and a head finding attention mechanism, the model created span embeddings. It was trained to optimize the marginal chance of gold antecedent spans from coreference clusters, allowing for possible mentions to be pruned.

The model learned a distribution $P(\cdot)$ over potential antecedent spans Y for each mention span x:

$$P(y) = \frac{e^{s(x,y)}}{\sum_{y' \in Y} e^{s(x,y')}}$$

The scoring function $s(x, y)$ between spans x and y represented its inputs using fixed-length span representations, $\mathbf{g_x}$ and $\mathbf{g_y}$. It computed the final score $s(x, y)$ by summing up three scores as shown below:

$$\begin{aligned} s(x,y) &= s_m(x) + s_m(y) + s_c(x,y) \\ s_m(x) &= \mathrm{FFNN}_m\left(\mathbf{g_x}\right) \\ s_m(y) &= \mathrm{FFNN}_m\left(\mathbf{g_y}\right) \\ s_c(x,y) &= \mathrm{FFNN}_c\left(\mathbf{g_x}, \mathbf{g_y}, \phi(x,y)\right) \end{aligned}$$

where the mention score $s_m(x)$ represents how likely span x is a mention, mention score $s_m(y)$ represents how likely span y is a mention, $s_c(x, y)$ represents how likely both x and y refer to the same entity assuming they are both mentions. FFNN$(\cdot)$ represents a feedforward neural network. Speaker and metadata characteristics are represented by $\phi(x, y)$.

The scoring method proposed by this work was an end-to-end neural network that computed the aforementioned scores given the document and its metadata. Vector representations $\mathbf{g_i}$ for each conceivable span i were created using bidirectional LSTMs (see Fig 1). They were made up of three vectors: the two BiLSTM hidden states (forward and backward) of the span endpoints, as well as an attention vector calculated over the span tokens. Instead of relying on syntactic parses, the model of Lee et al. (2017) developed a task-specific notion of headedness via an attention mechanism across words in each span.

$$\begin{aligned}
\alpha_t &= w_\alpha \cdot \mathrm{FFNN}_\alpha\left(\boldsymbol{x}_t^*\right) \\
a_{i,t} &= \frac{\exp\left(\alpha_t\right)}{\sum_{k=\mathrm{START}(i)}^{\mathrm{END}(i)} \exp\left(\alpha_k\right)} \\
\hat{\boldsymbol{x}}_i &= \sum_{t=\mathrm{START}(i)}^{\mathrm{END}(i)} a_{i,t} \cdot \boldsymbol{x}_t
\end{aligned}$$

where $\hat{x}_i$ denotes the weighted sum of word vectors in span i. x_t^* represents the vector representation for t-th word after going through the bidirectional LSTM. START(i) and END(i) denotes the indices for the starting and ending word of span i. $\boldsymbol{x}_t$ is the original word representation which is a concatenation of the works from Pennington et al. (2014), Turian et al. (2010) and Zhang et al. (2015). The weights $a_{i,t}$ were learned automatically and had a strong correlation with standard definitions of head words.

During the training phase, only clustering information is seen. The model maximized the marginal log-likelihood of all right antecedents suggested by gold clustering:

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

where GOLD(i) denotes the gold cluster that contains span i.

By maximizing this goal, the model learned to trim spans properly. While the first trimming was entirely arbitrary, only gold mentions received positive upgrades. The model could rapidly utilize this learning signal to award appropriate credit to the various parameters, such as the mention scores s_m used for trimming. This end-to-end model ensemble outperformed prior systems on the OntoNotes benchmark without the need of additional preprocessing techniques. This model learned to create useful mention candidates from the space of all possible spans implicitly. In addition, a unique headfinding attention mechanism learned a task-specific preference for head words.

7.4 Higher-order CR with coarse-to-fine inference

To encourage the coreference model to consider the cluster information from a global perspective, Lee et al. (2018) enhanced the end-to-end neural coreference model (Lee et al. 2017) via a novel approximation to higher-order inference. Lee et al. (2018) iteratively enhanced span representations with an attention mechanism. This allows the model to evaluate many hops in the anticipated clusters gently. Many previous coreference resolution methods have to rely on first order models (Clark and Manning 2016; Lee et al. 2017), which score only pairs of entity mentions. These models are computationally efficient and can handle large amounts of data. Since the models make individual decisions about coreference connections, they are subject to predicting clusters that are locally consistent but globally inconsistent. Lee et al. (2018) provided an iterative higher-order inference approximation based on a span-ranking architecture (Lee et al. 2017). At each iteration, the antecedent distribution was used as a focus mechanism to update current span representations, allowing upcoming coreference decisions to softly condition on past coreference decisions. The anticipated antecedent representation $\boldsymbol{a}_i^n$ of each span i at the n-th iteration was computed by using the current antecedent distribution $P_n\left(y_i\right)$, as shown below:

$$\boldsymbol{a}_i^n = \sum_{y_i \in \mathcal{Y}(i)} P_n\left(y_i\right) \cdot \boldsymbol{g}_{y_i}^n,$$

where $g_{y_i}^n$ denotes the vector representation of antecedent y_i of span i at the n-th iteration. Interpolation was then used to update the current span representation g_i^n with the predicted antecedent representation $\boldsymbol{a}_i^n$:

$$\begin{aligned} \boldsymbol{f}_i^n &= \sigma\left(\mathbf{W}_{\mathrm{f}}\left[\boldsymbol{g}_i^n, \boldsymbol{a}_i^n\right]\right) \\ \boldsymbol{g}_i^{n+1} &= \boldsymbol{f}_i^n \circ \boldsymbol{g}_i^n + \left(\mathbf{1} - \boldsymbol{f}_i^n\right) \circ \boldsymbol{a}_i^n \end{aligned}$$

where ∘ denotes element-wise multiplication. The learned vector $\boldsymbol{f}_i^n$ decides whether to maintain the existing span information or to integrate fresh information from its predicted antecedent for each dimension. $\boldsymbol{g}_i^n$ can be regarded as an element-wise weighted average of roughly n span representations.

Lee et al. (2018) additionally provided a coarse-to-fine approach that was learned with a single end-to-end goal to alleviate the computational issues associated with this higher-order inference. In the paired scoring function, they incorporated a coarse factor that was less accurate but more efficient. This additional factor allows another pruning step decreasing the number of antecedents to be evaluated. Before employing a more expensive scoring technique, the method intuitively computed a simple sketch of probable antecedents with a bilinear scoring function by

$$s_{\mathrm{c}}(i,j) = \boldsymbol{g}_i^{\top} \mathbf{W}_{\mathrm{c}} \boldsymbol{g}_j,$$

where $\mathbf{W}_{\mathrm{c}}$ denotes a learned weight matrix. The bilinear $s_{\mathrm{c}}(i,j)$ is less accurate but considerably fast. The pairwise score was then enhanced as follows:

$$s(i,j) = s_{\mathrm{m}}(i) + s_{\mathrm{m}}(j) + s_{\mathrm{c}}(i,j) + s_{\mathrm{a}}(i,j),$$

where $s_{\mathrm{m}}(i)$ denotes the score span i being a mention. $s_{\mathrm{m}}(j)$ denotes the score span j being a mention. $s_{\mathrm{c}}(i,j)$ denotes the score span i and span j being coreferent with less accuracy. $s_{\mathrm{a}}(i,j)$ denotes high accuracy coreference score of span i and span j. A three-stage beam search was used in the inference process: First, save the top M spans depending on each span mention score $s_{\mathrm{m}}(i)$. Secondly, keep the top K antecedents of each remaining span i, depending on the top three factors except $s_{\mathrm{a}}(i,j)$. Finally the overall coreference $s(i, j)$ is computed.

Although Lee et al. (2018) improved the performance of end-to-end model (Lee et al. 2017) via higher-order inference, at a later stage of study, Xu and Choi (2020) showed that the higher order inference concept has no direct impact on the performance improvement if the coreference resolution model uses more advanced encoder such as SpanBERT (Joshi et al. 2020).

8 Knowledge-based models

Similar to neural-based contextual models, recent knowledge-based models are likely trained with neural networks. However, besides the training datasets, knowledge-based models also explicitly employ external knowledge (e.g. general commonsense knowledge base or domain-specific knowledge base) which are usually stored in triplets.

8.1 Rewarding coreference resolver with world knowledge

Aralikatte et al. (2019) integrated knowledge information from Wikipedia and Wikidata in reinforcement learning models, taking into account coreference resolution methods that employed world knowledge. They improved performance of coreference resolvers by submitting predictions to an OpenRE system, comparing obtained relations with a knowledge base. The model illustrated that comparing the produced connections to the knowledge base was an indirect indicator of the quality of the CR. To generalize beyond the knowledge base, a Universal Schema model (Riedel et al. 2013) was developed and its confidence was utilized as the reward function. A policy-gradient fine-tuning of the coreference resolver was done with this incentive function, successfully maximizing the congruence of the predictions with world knowledge. An open information retrieval system (Angeli et al. 2015) converted each resolved document into n subject-relation-object (SRO) triples to update the coreference resolver. There was a reward function applied on each triple t_i to obtain a reward r_i for it. At the document level, the total reward was the normalized sum of the individual ones by

$$R = \frac{\sum_i r_i - \text{mean}\left(R_h\right)}{\text{stddev}\left(R_h\right)}$$

Where R_h is a sliding window containing the past $h = 100$ values. Policy gradient training was employed to update the coreference resolver because R was not differentiable in terms of the parameters of the coreference resolver. Exploration (choosing random actions) is used in addition to exploitation (sampling from the top actions predicted by the model) when choosing the optimal action.

In order to model consistency with world knowledge, several Universal Schema models (Riedel et al. 2013; Verga and McCallum 2016) were trained, resulting in three reward functions that predicted whether two items were connected, co-occur, or both connected and co-occur in Wikidata, respectively. The three incentives were focused on three different aspects of entity relations, providing complementary perspectives on how entities were connected. RE-Distill was generated by interpolating the learned weights of three reward models. As a result of this, the model provided three different policies: Coref-KG, Coref-Text, and Coref-Joint, which were trained by supervised learning and fine-tuned by three reward functions, respectively. The model then integrated these three strategies through multi-task reinforcement learning. This method was based on DisTraL (Teh et al. 2017), using policy gradients and model interpolation. Finally, the combined strategy was fine-tuned by the combined reward function. According to Aralikatte et al. (2019) the top performing fine-tune system outperformed the model from Lee et al. (2018) in terms of mention identification and linkage.

8.2 A generalized knowledge hunting framework

Emami et al. (2018) developed an automated system that excelled in Winograd Schema Challenge (WSC) and the Choice of Plausible Alternatives (COPA) tasks. The problem instances of these tasks need various, complicated types of reasoning and knowledge to solve. To gather texts from the web as evidences for possible issue resolutions, they used a knowledge-hunting module in their approach. Their approach generated suitable search engine queries depending on an input issue. It gathered and categorized information from returned results, and making decisions.

The knowledge hunting framework took a Winograd phrase as input and processed it through three steps before arriving at the final coreference determination. It started by mapping the phrase to a semantic representation schema, and then generating a sequence of queries that encapsulated the predicates of clauses. After passing the query set to a search engine, the search engine produced text results that fitted the schema. Once all of the returned snippets had been resolved, the results were used to create an estimate as to how the original Winograd problem would be addressed.

As part of their query generating process, Emami et al. (2018) utilized Stanford CoreNLP coreference resolver (Raghunathan et al. 2010b) to discover the predicates from the syntactic parse, as well as to extract the coreference chain of a potential evidence sentence during antecedent selection. For web scraping, Python Selenium module was utilized, and the search results were top two pages from Bing-USA and Google respectively. Each of the search results was comprised of an assortment of document snippets that included the search query. The framework then extracted the sentence(s) containing the query phrases, with the additional constraint that the words must be no more than 70 characters apart to guarantee relevancy.

8.3 Knowledge-aware pronoun CR

Zhang et al. (2019) investigated how to use external knowledge and contextual information to improve coreference resolution. To ensure the model generalizability, they incorporated information directly in the form of triplets Lin et al. (2023), which is the most common format for contemporary knowledge graphs, rather than features or rules, as is the case with traditional methods. Additionally, the authors proposed a knowledge attention module that improved their model by learning to choose and apply useful information depending on circumstances. The validity and effectiveness of the model were shown experimentally on two different datasets, where the proposed model outperformed baselines by a large margin. Additionally, since their model learned to include external input in addition to training data, it beat baselines in cross-domain situations.

The main architecture of the model was made up of numerous levels. To integrate contextual information, all mention spans (s) and pronouns (p) were embedded at the bottom. In the intermediate layer, the model utilized the embeddings of each pair of (s, p) to pick the most useful knowledge triplets from the knowledge base $\mathcal{G}$ and created the knowledge representation. The model concatenated the textual and knowledge representations and then predicted whether they had a coreference connection.

To encode each span, a conventional bidirectional LSTM (BiLSTM) (Graves and Schmidhuber 2005) was employed in this study. Various approaches were used to extract knowledge from a KG for pronouns and antecedents. The string match was utilized in the model for information extraction for the sake of simplicity and generality. In particular, for each triplet t in $\mathcal{G}$, the algorithm considered it to be a related triplet if its head was the same as the string of s. As a consequence, by averaging the embeddings of tail words, the model encoded t's information. The final score was as follows:

$$F(s,p) = f_m(s) + f_c(s,p)$$

where $f_m(s) = FFNN_m\left(\left[\mathbf{e}_s, \mathbf{o}_s\right]\right)$ is the function to determine if s is a valid mention and $f_c(s,p) = FFNN_c\left(\left[\mathbf{e}_s, \mathbf{o}_s, \mathbf{e}_p, \mathbf{o}_p, \mathbf{e}_s \odot \mathbf{e}_p, \mathbf{o}_s \odot \mathbf{o}_p\right]\right)$ is the function to represent the coreference strength between p and s, with $\odot$ representing multiplication by elements. $FFNN_m$ is feed forward neural network to calculate the mention score, $FFNN_c$ is the feed forward neural network to calculate the coreference score. $\mathbf{e}$ and $\mathbf{o}$ are span and knowledge representations respectively. Their subscripts s and p represents the mention span and pronoun, respectively. Following the calculation of the coreference score for all mention spans, The model selected the best candidate with a softmax function, which is defined as

$$\hat{F}(s,p) = \frac{e^{F(s,p)}}{\sum_{s_i \in \mathcal{S}} e^{F\left(s_i,p\right)}}.$$

The studies used two datasets: CoNLL and i2b2. Commonsense knowledge graph, medical ideas, linguist characteristics, and selectional preference were examples of knowledge resources. The author compared their model against the model of Lee et al. (2018). Experiment findings showed that their model had a more prominent performance under a cross-domain situation.

9 Transformer-based pre-trained models

Prior to the introduction of transformer architecture, the most widely used sequence conversion models were built on top of advanced CNN or RNN models that included both encoding and decoding processes. For boosting performance, the best versions included attention mechanisms to link the encoder and decoder together. Vaswani et al. (2017) proposed Transformer, a new basic network designed based on attention mechanism, without recurrent or convolutional structures. Transformer is trained considerably quicker than recurrent- or convolutional-based encoders for seq2seq tasks, such as language translation.

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019) is a Transformer-based pre-trained language model. BERT is designed to pre-train representations from unlabeled texts by conditioning all layers on the entire sentence. As a consequence, BERT can be fine-tuned with just one additional layer to offer cutting-edge models for a broad variety of tasks, such as question answering and language inference, without needing substantial architecture modifications. In this section, we introduce coreference resolution models that incorporate contextualized representations from BERT and its variants.

9.1 BERT for coreference resolution

Joshi et al. (2019) used BERT to resolve coreferences. Their model is based on top of a coarse to fine coreference model, termed c2f-coref for short from Lee et al. (2018).

In the work of Joshi et al. (2019), BERT fully replaced the LSTM-based encoder in c2f-coref. The span representations were the concatenation of beginning word piece, ending word piece, as well as the attended form of the whole span. Joshi et al. (2019) divided documents into parts before applying BERT in two ways: independent method and overlap method. The independent variation employed independent segments without overlapping. Each segment served as an independent input of BERT. By moving a sliding window T/2 steps each time, the overlap variations split the document into T-sized overlapping sections. The BERT encoder was then fed each segment separately, and the final representation was created via element-wise interpolation of the overlapping segment embeddings. The BERT-based models were tested against two datasets: the paragraph-level GAP dataset (Webster et al. 2018), and the document level CoNLL 2012 dataset (Pradhan et al. 2012). According to Joshi et al. (2019), both BERT-base and BERT-large outperformed the ELMo-based c2f-coref, with BERT-large beating the original c2f-coref by a larger margin.

9.2 SpanBERT for coreference resolution

Joshi et al. (2020) proposed SpanBERT pre-trained language model, developing a coreference resolution task-specific model by combining SpanBERT and the work of Lee et al. (2018). The main difference between Joshi et al. (2020) and Joshi et al. (2019) is that Joshi et al. (2020) used SpanBERT as the encoder rather than BERT. SpanBERT has the same architecture as BERT (Devlin et al. 2019), whereas it was trained with a different masking method in which spans were masked. The outer boundaries of spans were trained to predict all tokens inside the masked spans, which was called span-boundary objective (SBO). It is useful for coreference resolution, since entity mentions are often spans of tokens. Span ranking models benefit from better span representations.

SpanBERT retained the regular masked language model (MLM) objective in vanilla BERT but substituted SBO for next sentence prediction (NSP), because Joshi et al. (2020) discovered that single sequence training outperformed bi-sequence training on downstream tasks.

In another research later, Xia et al. (2020) reduced the memory usage of the original SpanBERT (Joshi et al. 2020) with an incremental algorithm. It kept the track of clusters, each of which had its own representation. The model suggested a possible set of spans for a particular phrase or segment. A scorer compared each span representation to all of the clusters, determining the the best fit cluster. Following the addition of the new span, the representation of the chosen cluster was likewise changed. The model periodically evicted less important entities and wrote them to disk. Each clustering choice made by this method was permanent.

Lai et al. (2022) incorporated the SpanBERT encoder into the e2e-coref model of Lee et al. (2017) but introduced a few simplifications to the original e2e-coref structure. Lai et al. (2022) excluded span length information when generating span representation and excluded feature information such as genre and distance when doing mention linking. It also reduced the number of candidate mentions when doing mention extraction. Despite those simplifications, Lai et al. (2022) still achieved comparative results with Joshi et al. (2020).

9.3 CorefQA

Unlike most previously discussed methods that have no chance of recovering a missed mention, Wu et al. (2020) permitted the mention linking module to find mentions that were missed during the mention proposal phase. The proposed model CorefQA defined coreference resolution as a span prediction issue under a question answering setting. It first generated a query for each mention before extracting the relevant mentions depending on the query. As long as at least one of the candidates in the associated coreference cluster was utilized in the query, other candidates in the cluster could be recovered during the mention linking phase.

CorefQA was made up of two modules: mention proposal and mention linking. The mention score was calculated using a FFNN taking into account the SpanBERT representation of the first and last constituent token of spans. Only spans with a mention score greater than a predefined threshold were retained in this module.

For the mention linking module, the query and the context were combined into a single sequence, and BIO (Beginning, Inside and Outside) tags were assigned to tokens that constituted candidate mentions. The probability of assigning one of BIO tags to a certain token was calculated via feed forward neural network. Thus, the probability of span j being coreferent to i depended on the probability that BIO tags were assigned correctly.

Furthermore, Wu et al. (2020) augmented data with Quoref dataset (Dasigi et al. 2019) and the SQuAD dataset (Rajpurkar et al. 2016), as well as using the speaker modeling strategy which directly combined the speaker names with the utterance, rather than converting the speaker information into binary features.

9.4 Using type information to improve entity coreference resolution

Khosla and Rose (2020) incorporated semantic knowledge into the model of Bamman et al. (2020). It reduced errors that were caused by type mismatches in coreference resolution. For each token, the model of Khosla and Rose (2020) first passed the BERT embeddings through a bi-directional LSTM in order to get the corresponding representation. The representation of a mention was given by a concatenation of token representations and different features including entity type. Coreference score of two mentions was given by a feedforward neural network whose input is the concatenation of mention representations, their element wise product and different mention-pair features including whether they have identical entity types. Empirical result showed that models incorporating type information outperformed baseline models without type information on four coreference resolution datasets. Thus Khosla and Rose (2020) argued that explicitly incorporating external knowledge would further benefit contextualized embedding-based models, e.g., BERT-based models.

9.5 Reinforcement learning based neural CR system

Wang et al. (2021) introduced a reinforcement learning-based resolver capable of handling problems, caused by the same mentions appearing in different document contexts. They utilized mention-level training examples, rather than merely sentence- or document-level samples. This algorithm has the advantage of mitigating the detrimental effect of noisy sentence-level information while retaining enough contextual information. The distance between two mentions was also taken into account in the work of Wang et al. (2021), since co-reference is sensitive to the mention distance. The span representation of Wang et al. (2021) was a combination of BERT embedding and a head-finding attention mechanism. The representations of two spans to be judged were then passed on to a actor-critic-based reinforcement learning model with two neural networks representing actor and critic separately. The states were the concatenation of the two mention spans. Action was defined as whether to create and store the links between the two spans and then move on to the next pair of spans. Reward was a biaffine attention mechanism to model the probability for the two spans to be coreferential. It also considered the distance between the two spans as there was usually a inverse relation between the distance and the coreference probability.

9.6 BERT fine-tuned with WikiCREM

Kocijan et al. (2019) fine-tuned BERT with WikiCREM. When the model was trained, sentences containing one masked personal name and two candidate mentions were given and the goal was to choose the more suitable candidate from the two. The objective function was a combination of the negative log-likelihood of the correct candidate as well as the max-margin loss term of the two candidates. It was observed that this combination of losses consistently outperformed single loss terms alone on various tasks.

$$\begin{aligned}\mathcal{L} = & -\log \mathbb{P}(\mathbf{a} \mid S) + \\ & + \alpha \cdot \max(0, \log \mathbb{P}(\mathbf{b} \mid S) - \log \mathbb{P}(\mathbf{a} \mid S) + \beta)\end{aligned}$$

where **a** represents the correct candidate. **b** represents the incorrect candidate. S denotes the sentence that contains the masked personal name. α and β are hyperparameters that control the influences of the loss components.

9.7 Gender resolution by evidence pooling

Attree (2019) presented a evidence-based deep learning model for the GAP shared task. It includes two main components: Pronoun BERT module and Evidence Pooling module. Pronoun BERT module extracted the last layer embedding for the pronoun from the BERT model. Evidence Pooling module combined the clustering information from four other coreference resolution models: AllenNLP (Gardner et al. 2017), NeuralCoref,[2] Parallelism+URL (Webster et al. 2018) and e2e-coref (Lee et al. 2017). The Evidence Pooling would encode the information from all these models via self-attention mechanism and generate an evidence vector. The readers is referred to Attree (2019) for details about how this evidence vector is generated. Finally, the evidence vector is concatenated with the BERT embedding of pronoun to go through the linear and softmax layer to get the classification result.

9.8 Other transformer-based CR models

CorefBERT(Ye et al. 2020) employed two training tasks: mention reference prediction (MRP) and MLM. For the input tokens $X = \left(x_1, x_2, \ldots, x_n\right)$, each token was first

[2] https://github.com/huggingface/neuralcoref.

represented by aggregating the embeddings of token and positional information, and then the input representations were fed into the bidirectional Transformer to obtain hidden states $\boldsymbol{H} = \left(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n\right)$, which were then used to compute the loss. The final loss function of CorefBERT was the sum of MRP loss and MLM loss, among which, the MRP loss was defined as a function to jointly maximize all the probability of choosing a word in the sequence to recover the masked word.

Yu et al. (2020) presented pairwise representation learning (PAIRWISERL) which was used for both entity coreference resolution and event coreference resolution. It treated entity coreference as a simplified version of event coreference resolution because event coreference resolution also includes arguments besides trigger itself. When processing event coreference resolution, PAIRWISERL concatenated two sentences containing the two events and passed it through RoBERTa (Liu et al. 2019) to get the representation for event triggers and four arguments: subject, object, time, and location. For each argument, PAIRWISERL concatenated representation from both sentences as well as their element wise product, then passing the concatenated vector through feed-forward neural networks in order to get the compatibility scores for that argument. The final binary classification result was given by a multilayer perceptron where the inputs is the concatenation of RoBERTA representation of two trigger words, their element wise product and the compatibility scores for the four arguments.

Lai et al. (2021) proposed a gating mechanism to selectively extract information from predicted features. The predicted features of event mentions included type, polarity, modality, genericity, tense and realis (Mitamura et al. 2016). For each event mention, its own representation was obtained using SpanBERT encoder (Joshi et al. 2020). The K symbolic features were converted into K vectors, using trainable matrices. Lai et al. (2021) then proposed a context-dependent gated module to filter information for each feature. In addition, Lai et al. (2021) introduced the noisy training method for regularization by randomly replacing some predicted feature values with some noise before feeding the input data into the model. By doing this, it could force the model to identify reliable features.

Caciularu et al. (2021) presented Cross Document Language Modeling (CDLM). All the related documents were concatenated and fed to the Longformer encoder (Beltagy et al. 2020) during pre-training. Caciularu et al. (2021) masked 15% of the tokens in each training example and forced the model to predict the masked token, based on the whole set of documents, rather than the individual document. Caciularu et al. (2021) employed the Multi-News dataset (Fabbri et al. 2019) which contains 44972 document clusters for pre-training. For each pre-training example, documents within the same cluster were randomly picked in order to make sure that the documents were related. During the fine-tuning of coreference resolution, relevant documents were concatenated into a single sequence with document separator tokens ([CLS]) at the beginning of the sequence. The pair-wise vector representation $m_t(i,j)$ between mention i and mention j within the t-th example was the concatenation of CDLM representations of the [CLS] token, mention i and j, and their element wise product. $m_t(i,j)$ was then passed through a multi-layer perceptron to get the binary classification result (coreferent or not).

In order to reduce the large memory footprint faced by many previous models, Kirstain et al. (2021) presented start-to-end (s2e) model which only use information on the start and end tokens of the span in order to calculate the mention score and antecedent score. By doing this, it reduced the memory footprint significantly compared with Joshi et al. (2020). s2e model utilized bilinear functions between pairs of endpoints tokens to calculate mention score f_m and antecedent score f_a without relying on the span level representation.

Similar to Kirstain et al. (2021), Thirukovalluru et al. (2021) also aimed at reducing the memory and time cost of coreference resolution systems. They presented an approximation to the end-to-end model Lee et al. (2017) which can scale to long documents. Beside using token level bilinear inference to calculate scores, it also proposed other tricks such as token k-nearest neighbour approximation, an approximation to the token similarity matrix and also a probing approach to drop less important tokens.

Cattan et al. (2021) presented an end-to-end model that focus on cross document (CD) coreference resolution. It first pre-trained the mention scorer $s_m(\cdot)$ on the gold mention spans of ECB+ dataset. During the training phase, the pairwise scorer $s_a(i, j)$ compared the mention with all the spans across all the documents and optimized the cross entropy loss of mention-pair scores.

In order to identify paraphrase relations between event mentions and avoid the propagation errors, Zeng et al. (2020) proposed Event-specific Paraphrases and Argument-aware Semantic Embeddings (EPASE). EPASE improved generalization ability in two aspects: recognizing event paraphrases under more situations and incorporating the argument roles into the event mention embedding.

Yadav et al. (2021) proposed a way to solve event and entity coreference resolution jointly under the cross-document coreference resolution. It took the uncertainty of coreference decision into consideration when defining the cost function. The joint coreference model built cluster trees to represent the uncertainty with mentions as its leaves and trained a joint cost function. The core idea of the joint cost function relied on two parts: pairwise mention scorer and relational similarity. Pairwise mention score was calculated via the model proposed by Cattan et al. (2020). The mention pair's RoBERTa encoded representation and their element wise product were concatenated and passed through an MLP to get the score. As for relational similarity, it was a weighted average of the similarity score of different arguments of the event mentions. The similarity score of the arguments was calculated based on different properties of the structure of the cluster tree.

Another example of joint learning in coreference resolution is Lu and Ng (2021a) in which the models jointly learned six related tasks: trigger detection, entity coreference, anaphoricity detection, realis detection, argument extraction, and event coreference. The model also used consistency constraints to guide this multi-task learning process. Lu and Ng (2021b) further did an empirical analysis of this model and draw a few interesting findings such as event CR performance could be enhanced by improving mention boundary detection, anaphoricity detection, and subtype detection.

Dobrovolskii (2021) proposed a word-level coreference resolution model wl-coref that focused on individual words rather than spans in order to reduce the complexity of model. It first constructed each word's representation by combining the constituent tokens' contextualized representation. Wl-coref used a bilinear function to get the most possible antecedents for each token. Then for each candidate antecedent, its coreference score was calculated by a feed-forward neural network taking into consideration token embeddings as well as feature information such as distance and speaker. Finally, a feature extraction module was employed to to determine the boundaries of spans based on word-level coreference links.

Table 1 The statistics and features for different coreference resolution datasets

Dataset	Size				Tasks	
	Train	Val	Test	Total	Entity	Event
CoNLL 2012	2802	343	348	3493	✓	
GAP	4000	908	4000	8908	✓	
KBP 2017[1]	–	–	167	167		✓
ACE 2005[2]	529	28	40	599		✓
LitBank	–	–	–	100	✓	
WSC	554	104	146	804	✓	
DPR	1322	–	564	1886	✓	
PDP	–	–	–	60	✓	
Winogender	–	–	–	720	✓	
WinoBias	1580	–	1580	3160	✓	
KnowRef	7455	–	1269	8724	✓	
WikiCoref	–	–	–	30	✓	
ECB+	574	196	206	976	✓	✓
RED	–	–	–	95		✓
GUM	–	–	–	168	✓	
WEC	40529	1250	1893	43672		✓
EmailCoref	36	–	10	46	✓	
BUG	–	–	108K	108K	✓	

[1] Some research (Huang et al. 2019; Lu et al. 2022; Yu et al. 2020) combined KBP 2015 (Ellis et al. 2015), KBP 2016 (Ellis et al. 2016) with KBP 2017, using KBP 2015 (360 documents) as the training set, KBP 2016 (169 documents) as the validation set and KBP 2017 (167 documents) as the test set

[2] This split is based on Lin et al. (2020)

10 Summary of datasets and models

10.1 Summary of datasets

Table 1 summarizes the statistics and features of different datasets used in the coreference resolution research field. The four columns under size category summarizes number of examples in training set, validation set, test set and whole dataset, respectively. The task columns presented whether the dataset is focusing on entity coreference resolution or event coreference resolution. We can observe from Table 1 that KBP 2017, ACE 2005, ECB+, RED and WEC are mainly used for event coreference resolution, where ECB+ can be used for entity coreference resolution as well. For the rest of the datasets, they are mainly used for entity coreference resolution research.

Table 2 summarizes the application scene of different datasets under the "aim" column. We have also listed the state of the art (SOTA) models for each dataset. The details of those models are discussed in the following model sections. It is important to note that not all the models were evaluated on the same datasets as some models are designed to address different challenges, thus, not all models are directly comparable to each other.

Table 2 The aim, state of the art (SOTA) models and their performances for different coreference resolution datasets

Dataset	Aim	SOTA model	Metrics and result for SOTA
CoNLL 2012	Shared task corpus	Wang et al. (2021)	CoNLL score: 87.5%
GAP	Gender bias in PCR	Attree (2019)	F1: 92.5% gender bias:0.97
KBP 2017	Within-document Event Coreference	Yu et al. (2020)	AVG-F: 57.12%
ACE 2005	Within-document Event Coreference	Lai et al. (2021)	CoNLL score: 87.9% AVG-F:88.30%
LitBank	Long-distance within-document coreference	Khosla and Rose (2020)	CoNLL score: 80.26%
WSC	Commonsense knowledge in PCR	Sun et al. (2021)	Accuracy: 97.3%
DPR	Complex cases of definite pronouns	Kocijan et al. (2019)	Accuracy: 84.8%
PDP	Commonsense knowledge in PCR	Kocijan et al. (2019)	Accuracy: 86.7%
Winogender	Gender bias in PCR	Kocijan et al. (2019)	Accuracy: 82.1%
WinoBias	Occupational gender bias in PCR	Kocijan et al. (2019)	Subset1 accuracy: 78.0%–78.2% Subset2 accuracy: 98.7%–99.0%
KnowRef	Challenging cases in PCR	Emami et al. (2019)	Accuracy: 71%
WikiCoref	Coreference on Wikipedia	Khosla and Rose (2020)	CoNLL score: 71.35%
ECB+	Cross-document Coreference	Caciularu et al. (2021)	CoNLL score: 85.6%
RED	Within-document Event Coreference	–	–
GUM	Shared task corpus	–	–
WEC	Cross-document Event Coreference	Eirew et al. (2021)	CoNLL score: 62.3%
EmailCoref	Entity coreference in email threads	Khosla and Rose (2020)	CoNLL score: 76.17%
BUG	Gender bias in PCR	Levy et al. (2021)	Accuracy: 64.1%

PCR denotes pronoun coreference resolution

Table 3 The tasks and learning techniques of different models

Models	Target tasks		Rule-based	Traditional machine learning			Deep learning			
	Entity	Event	Method	DT	MBL	CRF	GM	RNN	AM	RL
Ding and Liu (2010)	✓			✓						
Atkinson et al. (2015)	✓				✓					
Durrett and Klein (2014)	✓					✓				✓
Clark and Manning (2015)	✓									
Wiseman et al. (2016)	✓							✓		
Lee et al. (2017)	✓							✓	✓	
Lee et al. (2018)	✓						✓	✓	✓	
Aralikatte et al. (2019)	✓							✓		✓
Emami et al. (2018)	✓		✓							
Zhang et al. (2019)	✓							✓	✓	
Joshi et al. (2019)	✓								✓	
Joshi et al. (2020)	✓								✓	
Kocijan et al. (2019)	✓									
Ye et al. (2020)	✓									
Wu et al. (2020)	✓								✓	
Khosla and Rose (2020)	✓							✓	✓	
Wang et al. (2021)	✓								✓	✓
Attree (2019)	✓								✓	
Yu et al. (2020)	✓	✓								
Lai et al. (2021)		✓					✓			
Caciularu et al. (2021)	✓	✓								
Kirstain et al. (2021)	✓									
Thirukovalluru et al. (2021)	✓									
Cattan et al. (2021)	✓	✓								
Zeng et al. (2020)		✓								
Yadav et al. (2021)	✓	✓								
Dobrovolskii (2021)	✓									

DT denotes decision tree-based method, *MBL* denotes memory-based learning, *CRF* denotes conditional random field, *GM* denotes gated mechanism, *AM* denotes attention mechanism, *RL* denotes reinforcement Learning

10.2 Tasks and learning methods

Table 3 summarizes the target tasks and learning methods of the models from Sect. 6 to Sect. 9. As can be seen in Table 3, target tasks columns specify whether the model mainly aims at solving entity coreference resolution or event coreference resolution problems. Learning methods specify whether the model employed a rule-based method, traditional machine learning-based methods, or deep learning-based methods. Models presented in Sect. 6 are mainly based on traditional machine learning methods, whereas models in Sects. 7–9 are mainly based on deep learning. Among the deep learning-based models,

Table 4 The features employed by different models. SC denotes semantic consistency, OW denotes opinion words. TC denotes text chunking, WP denotes word position

Models	Semantic features			Syntactic features			Word	Pre-trained
	SC	OW	NER tags	POS tags	TC	WP	Embedding[a]	LM[b]
Ding and Liu (2010)	✓	✓		✓	✓			
Atkinson et al. (2015)			✓	✓				
Durrett and Klein (2014)						✓		
Clark and Manning (2015)			✓	✓		✓		
Wiseman et al. (2016)			✓	✓		✓		
Lee et al. (2017)							G,T	
Lee et al. (2018)							G,E,T	
Aralikatte et al. (2019)							G,E,T	
Emami et al. (2018)								
Zhang et al. (2019)							G,E	
Joshi et al. (2019)								B
Joshi et al. (2020)								S
Kocijan et al. (2019)								B
Ye et al. (2020)								C
Wu et al. (2020)								S
Khosla and Rose (2020)			✓					B
Wang et al. (2021)						✓		B
Attree (2019)								B
Yu et al. (2020)								R
Lai et al. (2021)								S
Caciularu et al. (2021)								L
Kirstain et al. (2021)								L
Thirukovalluru et al. (2021)								S
Cattan et al. (2021)								R
Zeng et al. (2020)								B
Yadav et al. (2021)								R
Dobrovolskii (2021)								R

[a] In the word embedding column, *G* denotes GloVe embedding, *E* denotes ELMo embedding, *T* denotes Turian embedding

[b] In the pre-trained language model (LM) column, *B* denotes BERT, *S* denotes SpanBERT, *C* denotes Coref-BERT, *R* denotes RoBERTa, *L* denotes LongFormer

contextual-based models and knowledge-based models mainly employ recurrent neural network as their basic structure, except for the works of Clark and Manning (2015) and Emami et al. (2018). For large scale pre-trained language model-based methods, recurrent neural network encoders have been replaced with Transformers. Additional special techniques, such as gated mechanism, attention mechanism, and reinforcement learning are also employed by some deep learning-based models. E.g., Lai et al. (2021) employed gated mechanism to filter feature information for event coreference resolution. Wang et al. (2021) employed an attention layer on top of BERT embedding layer in order to assign different

Table 5 External knowledge used in knowledge-based models

Models	Wikipedia	Wikidata	SE	OMCS	MC	PAG
Aralikatte et al. (2019)	✓	✓				
Emami et al. (2018)			✓			
Zhang et al. (2019)	✓			✓	✓	✓

SE denotes Search Engine, *OMCS* denotes open mind commonsense knowledge base, *MC* denotes medical concepts, *PAG* denotes plurality, animacy & gender

weights to different tokens within the same mention. Aralikatte et al. (2019) employed reinforcement learning to award model for being consistent with world knowledge.

10.3 Features

Table 4 presents the different features employed by the models described from Sect. 6 to Sect. 9. The features include semantic features, syntactic features, word embeddings and pre-trained language models. Syntactic and semantic features are mainly employed by feature-based models and some early stage neural network models before Lee et al. (2017). From Lee et al. (2017) onwards, mentions are mainly represented using word embeddings. From Joshi et al. (2019) onwards, mentions are mainly represented using pre-trained language models. Although it is uncommon to see pre-trained language model-based methods to employ semantic and syntactic information explicitly, there are two exceptions with Khosla and Rose (2020) employing NER tags and Wang et al. (2021) taking word position into consideration.

10.4 External knowledge

In Sect. 8, we have introduced representative models that explicitly employ external knowledge during coreference resolution. We summarize the knowledge used by those models in Table 5. Aralikatte et al. (2019) employed knowledge triplets from Wikipedia and Wikidata in order to compare the predicted connection with knowledge base for reward generation purpose. Emami et al. (2018) used search engine to gather texts from web for an evidence-based coreference reasoning for WSC examples. Zhang et al. (2019) employed knowledge from open mind commonsense[3] (OMCS), medical concepts (Uzuner et al. 2012), the linguistic features of plurality, animacy & gender and selectional preferences from Wikipedia in order to generate knowledge embeddings for candidate spans.

11 Challenges and future work

So far, we have shown four technical trends of coreference resolution models. Despite the fact that there are no absolute boundaries between their timelines, we can roughly conclude that over the last decade, the research interest has shifted from feature-based models that

[3] https://www.media.mit.edu/projects/open-mind-common-sense/overview/.

incorporate traditional machine learning to deep learning models that rely on contextual and explicit knowledge base information, and then to methods that are built on top of transformer-based large scale pre-trained models. According to Sect. 4, all the state-of-the-art models on the aforementioned datasets are based on large-scale pre-trained models. This shows the strength of this technical trend.

Coreference Resolution has come a long way with substantial improvements in recent years. However, it is clear that coreference resolution has been a difficult task, with problems still to be addressed in both research and practice.

11.1 Lack of datasets in downstream task with CR labels

CR is currently constrained by a shortage of resources as developing a dataset that shows CR contribution to downstream tasks and is of high quality is difficult and costly. Utilizing a consistent annotation method and procedure while developing these resources may be crucial to their final quality. Liu et al. (2021) have shown that coreference guidance improves conversation summarization, whereas the application of coreference resolution in other tasks such as dialogue systems should be explored further.

11.2 Lack of the combination of symbolic features with subsymbolic features

Majority of recent modeling techniques are based on Lee et al. (2017) and Joshi et al. (2019) and make extensive use of subsymbolic features. As pointed out by Khosla and Rose (2020), symbolic features like entity type could improve the subsymbolic coreference resolution systems' performance. To enhance the performance of CR systems in the future, combining additional symbolic features, such as semantic features and knowledge representations, and subsymbolic methods, such as word embeddings, may be investigated (Mao et al. 2018; Cambria et al. 2022; Mao et al. 2022).

11.3 Incorporating linguistic and cognitive intuition

There has been a wide range of linguistic and cognitive studies in analyzing different referring expression phenomena, e.g., discourse salience (Miltsakaki 2007), donkey sentences (Brasoveanu 2008), and pronoun-dropping (also termed pro-drop or null anaphora) (Bussmann et al. 2006). These research outcomes have built a theoretical foundation for data annotations and model design. However, these linguistic and cognitive findings were rarely incorporated in deep learning-based CR models. Deep learning models seem to fall into similar paradigms in various computational linguistics tasks. One may expect the deep learning-based CR models have more linguistic and cognitive intuition, because the findings of other NLP domains have shown that explicitly modeling the linguistic intuition, cognition and commonsense can further boost model performance in a specific task (Chaturvedi et al. 2019; Mao et al. 2019; Murugesan et al. 2021; Ge et al. 2022). Incorporating linguistic and cognitive intuition also helps deep learning models achieve human-like and explainable CR (Ellis et al. 2022).

11.4 Current models demand an excessive amount of resources

Current CR systems are highly resource-intensive in terms of memory and computing power, making them unsuitable for multitask learning with other modules. Novel methods have started to consider such limitations, attempting to minimize the memory requirements and computations required to construct coreference clusters. Efforts like knowledge distillation (Hinton et al. 2015) to condense the model could be potential answers to this issue.

11.5 Exploiting the advent of super large scale language models

The traditional pre-training and fine-tuning paradigm uses language models as the foundation and adds task-specific layers on top of it to fine tune the whole model. However, it would be quite impossible to incorporate the recently released super large scale language models such as ChatGPT (OpenAI 2022) or GPT-4 (OpenAI 2023) into this paradigm for two reasons: (1) this type of model would be extremely resource demanding, (2) these models are usually not open source to the public. How to exploit the advent of these super large scale language models to promote CR research might be an intriguing avenue to pursue.

12 Conclusion

Conversation and, by extension, language modeling and understanding rely heavily on coreference resolution. Despite substantial progress in recent years, it is still considered as one of the most challenging tasks in NLP due to the required commonsense and domain specific knowledge. This work seeks to provide a well-structured survey about recent advances in Coreference Resolution.

We present that coreference resolution has progressed over the past decade from feature-based techniques to contextual-based and external knowledge-based solutions. Additionally, it is shown that current state-of-the-art CR algorithms include large-scale pre-trained language models that implicitly contain both contextual and external information. Furthermore, we have provided a list of datasets and metrics that are critical for coreference resolution experiments. We believe that our study will aid scholars in the relevant area in establishing a solid foundation for coreference resolution development.

Acknowledgements This study is supported under the RIE2020 Industry Alignment Fund-Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

Abzaliev A (2019) On GAP coreference resolution shared task: insights from the 3rd place solution. In: Proceedings of the first workshop on gender bias in natural language processing, Florence, pp 107–112. Association for Computational Linguistics

Agarwal O, Subramanian S, Nenkova A, Roth D (2019) Evaluation of named entity coreference. In: Proceedings of the second workshop on computational models of reference, anaphora and coreference, Minneapolis, pp 1–7. Association for Computational Linguistics

Angeli G, Johnson Premkumar MJ, Manning CD (2015) Leveraging linguistic structure for open domain information extraction. 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, pp 344–354. Association for Computational Linguistics

Aralikatte R, Lent H, Gonzalez AV, Herschcovich D, Qiu C, Sandholm A, Ringaard M, Søgaard A (2019) Rewarding coreference resolvers for being consistent with world knowledge. 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), Hong Kong pp 1229–1235. Association for Computational Linguistics

Atkinson J, Salas G, Figueroa A (2015) Improving opinion retrieval in social media by combining features-based coreferencing and memory-based learning. Inform Sci 299:20–31. https://doi.org/10.1016/j.ins.2014.12.021

Attree S (2019), August. Gendered ambiguous pronouns shared task: Boosting model confidence by evidence pooling. In: Proceedings of the first workshop on gender bias in natural language processing, Florence, pp 134–146. Association for Computational Linguistics

Bagga A, Baldwin B (1998) Algorithms for scoring coreference chains. In: The first international conference on language resources and evaluation workshop on linguistics coreference, Volume 1, pp 563–566. Citeseer

Bamman D, Lewke O, Mansoor A (2020) An annotated dataset of coreference in English literature. In: Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, pp 44–54. European Language Resources Association

Beltagy I, Peters ME, Cohan A (2020) Longformer: the long-document transformer. *CoRR* abs/2004.05150. arXiv:2004.05150

Bhattacharjee S, Haque R, de Buy Wenniger GM, Way A (2020) Investigating query expansion and coreference resolution in question answering on BERT. In: International conference on applications of natural language to information systems, pp 47–59. Springer

Bornstein A, Cattan A, Dagan I (2020) CoRefi: a crowd sourcing suite for coreference annotation. In; Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations

Brasoveanu A (2008) Donkey pluralities: plural information states versus non-atomic individuals. Linguist Philos 31(2):129–209

Bussmann H, Kazzazi K, Trauth G (2006) Routledge dictionary of language and linguistics. Routledge, London

Caciularu A, Cohan A, Beltagy I, Peters M, Cattan A, Dagan I (2021) November. CDLM: cross-document language modeling. In: Findings of the association for computational linguistics: EMNLP 2021, Punta Cana, Dominican Republic, pp 2648–2662. Association for Computational Linguistics

Cambria E, Liu Q, Decherchi S, Xing F, Kwok K (2022) SenticNet 7: a commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In: Proceedings of the 13th language resources and evaluation conference, pp 3829–3839

Cattan A, Eirew A, Stanovsky G, Joshi M, Dagan I (2020) Streamlining cross-document coreference resolution: evaluation and modeling. *CoRR* abs/2009.11032. arXiv:2009.11032

Cattan A, Eirew A, Stanovsky G, Joshi M, Dagan I (2021) Cross-document coreference resolution over predicted mentions. In: Findings of the association for computational linguistics: ACL-IJCNLP 2021, Online, pp 5100–5107. Association for Computational Linguistics

Chaturvedi I, Satapathy R, Cavallari S, Cambria E (2019) Fuzzy commonsense reasoning for multimodal sentiment analysis. Pattern Recognit Lett 125:264–270

Chen G, Van DeemterK, Lin C (2018) Modelling pro-drop with the rational speech acts model. In: Proceedings of the 11th international conference on natural language generation, pp 57–66. Association for Computational Linguistics (ACL)

Clark K, Manning CD (2015) 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, pp 1405–1415. Association for Computational Linguistics

Clark K, Manning CD (2016) Deep reinforcement learning for mention-ranking coreference models. In: Proceedings of the 2016 conference on empirical methods in natural language processing, Austin, pp 2256–2262. Association for Computational Linguistics

Cybulska A , Vossen P (2014) Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In: Proceedings of the ninth international conference on language resources and evaluation (LREC'14), Reykjavik, pp 4545–4552. European Language Resources Association (ELRA)

Dai Z, Fei H, Li P (2019) Coreference aware representation learning for neural named entity recognition. In: Proceedings of the Twenty-eighth international joint conference on artificial intelligence, IJCAI-19, pp 4946–4953. International Joint Conferences on Artificial Intelligence Organization

Dakle PP, Desai T, Moldovan D (2020) A study on entity resolution for email conversations. In: Proceedings of the 12th language resources and evaluation conference, Marseille, pp 65–73. European Language Resources Association

Dasigi P, Liu NF, Marasović A, Smith NA, Gardner M (2019) QUOREF: a reading comprehension dataset with questions requiring coreferential reasoning. 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), Hong Kong, pp 5925–5932. Association for Computational Linguistics

Davis E, Morgenstern L, Ortiz C (2017) The first Winograd schema challenge at IJCAI-16. AI Mag 38(3):97–98. https://doi.org/10.1609/aimag.v38i4.2734

de Marneffe MC, Rafferty AN, Manning CD (2008) Finding contradictions in text. In: Proceedings of ACL-08: HLT, Columbus, pp 1039–1047. Association for Computational Linguistics

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and Short Papers), Minneapolis, pp 4171–4186. Association for Computational Linguistics

Ding X, Liu B (2010) Resolving object and attribute coreference in opinion mining. In: Proceedings of the 23rd international conference on computational linguistics (COLING 2010), Beijing, pp 268–276. COLING 2010 Organizing Committee

Dobrovolskii V (2021) Word-level coreference resolution. In: Proceedings of the 2021 conference on empirical methods in natural language processing, Online and Punta Cana, Dominican Republic, pp 7670–7675. Association for Computational Linguistics

Durrett G, Klein D (2014) A joint model for entity analysis: coreference, typing, and linking. Trans Assoc Comput Linguist 2:477–490. https://doi.org/10.1162/tacl_a_00197

Eirew A, Cattan A, Dagan I (2021) WEC: deriving a large-scale cross-document event coreference dataset from Wikipedia. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 2498–2510. Association for Computational Linguistics

Ellis J, Getman J, Fore D, Kuster N, Song Z, Bies A, Strassel SM (2015) Overview of linguistic resources for the TAC KBP 2015 evaluations: Methodologies and results. In: Proceedings of the 2015 text analysis conference, TAC 2015, Gaithersburg, November 16–17, 2015, 2015. NIST

Ellis J, Getman J, Kuster N, Song Z, Bies A, Strassel SM (2016) Overview of linguistic resources for the TAC KBP 2016 evaluations: Methodologies and results. In: Proceedings of the 2016 Text analysis conference, TAC 2016, Gaithersburg, November 14–15, 2016. NIST

Ellis K, Albright A, Solar-Lezama A, Tenenbaum JB, O'Donnell TJ (2022) Synthesizing theories of human language with Bayesian program induction. Nat Commun 13(1):1–13

Emami A, Trichelair P, Trischler A, Suleman K, Schulz H, Cheung JCK (2019) The KnowRef coreference corpus: removing gender and number cues for difficult pronominal anaphora resolution. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, pp. 3952–3961. Association for Computational Linguistics

Emami A, Trischler A, Suleman K, Cheung JCK (2018), June. A generalized knowledge hunting framework for the Winograd schema challenge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, New Orleans, Louisiana, USA, pp. 25–31. Association for Computational Linguistics

Fabbri A, Li I, She T, Li S, Radev D (2019) Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, pp 1074–1084. Association for Computational Linguistics

Ferracane E, Marshall I, Wallace BC, Erk K (2016) Leveraging coreference to identify arms in medical abstracts: an experimental study. In: Proceedings of the seventh international workshop on health text mining and information analysis, Auxtin, pp 86–95. Association for Computational Linguistics

Gardner M, Grus J, Neumann M, Tafjord O, Dasigi P, Liu NF, Peters M, Schmitz M, Zettlemoyer LS (2017) AllenNLP: a deep semantic natural language processing platform. In: Proceedings of workshop for NLP open source software (NLP-OSS)

Ge M, Mao R, Cambria E (2022) Explainable metaphor identification inspired by conceptual metaphor theory. In: Proceedings of the 36th AAAI conference on artificial intelligence, pp 10681–10689

Ghaddar A , Langlais P (2016) WikiCoref: an English coreference-annotated corpus of Wikipedia articles. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16), Portorož, Slovenia, pp 136–142. European Language Resources Association (ELRA)

Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 18(5):602–610. https://doi.org/10.1016/j.neunet.2005.06.042

Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. In: NIPS deep learning and representation learning workshop

Hovy E, Marcus M, Palmer M, Ramshaw L, Weischedel R (2006) OntoNotes: the 90% solution. In: Proceedings of the human language technology conference of the NAACL, companion volume: short papers, New York City, pp 57–60. Association for Computational Linguistics

Huang YJ, Lu J, Kurohashi S, Ng V (2019) Improving event coreference resolution by learning argument compatibility from unlabeled data. 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, pp 4171–4186. Association for Computational Linguistics

Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O (2020) SpanBERT: improving pre-training by representing and predicting spans. Trans Assoc Comput Linguist 8:64–77. https://doi.org/10.1162/tacl_a_00300

Joshi M, Levy O, Zettlemoyer L, Weld D (2019) 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), Hong Kong, pp 5803–5808. Association for Computational Linguistics

Khashabi D, Chaturvedi S, Roth M, Upadhyay S, Roth D (2018) Looking beyond the surface: a challenge set for reading comprehension over multiple sentences. 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, pp 252–262. Association for Computational Linguistics

Khosla S, Rose C (2020) Using type information to improve entity coreference resolution. In: Proceedings of the first workshop on computational approaches to discourse, pp 20–31. Association for Computational Linguistics

Kirstain Y, Ram O, Levy O (2021) Coreference resolution without span representations. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 2: Short Papers), pp 14–19. Association for Computational Linguistics

Kocijan V, Camburu OM, Cretu AM, Yordanov Y, Blunsom P, Lukasiewicz T (2019) WikiCREM: a large unsupervised corpus for coreference resolution. 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), Hong Kong, pp 4303–4312. Association for Computational Linguistics

Kopeć M (2014) MMAX2 for coreference annotation. In: Proceedings of the demonstrations at the 14th conference of the European chapter of the association for computational linguistics, Gothenburg, pp 93–96. Association for Computational Linguistics

Krishna MH, Rahamathulla K, Akbar A (2017) A feature based approach for sentiment analysis using SVM and coreference resolution. In: 2017 International conference on inventive communication and computational technologies (ICICCT), pp 397–399

Kuhn HW (1955) The Hungarian method for the assignment problem. Naval Res Logist Q 2(1–2):83–97

Kundu G, Sil A, Florian R, Hamza W (2018) Neural cross-lingual coreference resolution and its application to entity linking. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: short papers), Melbourne, pp 395–400. Association for Computational Linguistics

Lai T, Ji H, Bui T, Tran QH, Dernoncourt F, Chang W (2021) A context-dependent gated module for incorporating symbolic semantics into event coreference resolution. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 3491–3499. Association for Computational Linguistics

Lai TM, Bui T, Kim DS (2022) End-to-end neural coreference resolution revisited: a simple yet effective baseline. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 8147–8151

Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) ALBERT: a lite BERT for self-supervised learning of language representations. In: International conference on learning representations

Lee K, He L, Lewis M, Zettlemoyer L (2017) End-to-end neural coreference resolution. In: Proceedings of the 2017 conference on empirical methods in natural language processing, Copenhagen, pp. 188–197. Association for Computational Linguistics

Lee K, He L, Zettlemoyer L (2018) Higher-order coreference resolution with coarse-to-fine inference. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 2 (short papers), New Orleans, pp 687–692. Association for Computational Linguistics

Levesque H, Davis E, Morgenstern L (2012). The Winograd schema challenge. In: Thirteenth international conference on the principles of knowledge representation and reasoning. Citeseer

Levesque HJ (2011) The Winograd schema challenge. In: Logical formalizations of commonsense reasoning, Papers from the 2011 AAAI spring symposium, Technical Report SS-11-06, Stanford, March 21–23, 2011. AAAI

Levy S, Lazar K, Stanovsky G (2021) Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In: Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, pp 2470–2480. Association for Computational Linguistics

Li X, Van Deemter K, Lin C (2018) Statistical NLG for generating the content and form of referring expressions. In: Proceedings of the 11th international conference on natural language generation. Association for Computational Linguistics (ACL)

Lin Q, Mao R, Liu J, Xu F, Cambria E (2023) Fusing topology contexts and logical rules in language models for knowledge graph completion. Inform Fus 90:253–264

Lin Y, Ji H, Huang F, Wu L (2020) A joint neural model for information extraction with global features. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 7999–8009. Association for Computational Linguistics

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: a robustly optimized BERT pretraining approach. CoRR abs/1907.11692. arXiv:1907.11692

Liu Z, Shi K, Chen N (2021) Coreference-aware dialogue summarization. In: Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue, Singapore, pp. 509–519. Association for Computational Linguistics

Lu J, Ng V (2018) Event coreference resolution: a survey of two decades of research. In: Proceedings of the twenty-seventh international joint conference on artificial intelligence, IJCAI-18, pp 5479–5486. International Joint Conferences on Artificial Intelligence Organization

Lu, J, Ng V (2020) Conundrums in entity coreference resolution: Making sense of the state of the art. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp. 6620–6631. Association for Computational Linguistics

Lu J, Ng V, (2021a) Constrained multi-task learning for event coreference resolution. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 4504–4514. Association for Computational Linguistics

Lu J, Ng V, (2021b) Conundrums in event coreference resolution: Making sense of the state of the art. In: Proceedings of the 2021 conference on empirical methods in natural language processing, Punta Cana, pp 1368–1380. Association for Computational Linguistics

Lu J, Ng V. (2021c) Span-based event coreference resolution. In: Proceedings of the AAAI conference on artificial intelligence *35*(15): 13489–13497. https://doi.org/10.1609/aaai.v35i15.17591

Lu Y, Lin H, Tang J, Han X, Sun L (2022) End-to-end neural event coreference resolution. Artificial Intell 303:103632. https://doi.org/10.1016/j.artint.2021.103632

Luo X (2005) On coreference resolution performance metrics. In: Proceedings of human language technology conference and conference on empirical methods in natural language processing, Vancouver, pp 25–32. Association for Computational Linguistics

Luo X, Pradhan S (2016) Evaluation metrics, Anaphora resolution. Springer, Berlin, pp 141–163. https://doi.org/10.1007/978-3-662-47909-4_5

Mao R, Li X (2021) Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. Proc AAAI Conf Artif Intell 35(15):13534–13542

Mao R, Li X, Ge M, Cambria E (2022) MetaPro: a computational metaphor processing model for text pre-processing. Inform Fus 86–87:30–43

Mao R, Lin C, Guerin F (2018) Word embedding and WordNet based metaphor identification and interpretation. In: Proceedings of the 56th annual meeting of the association for computational linguistics, vol 1, pp 1222–1231

Mao R, Lin C, Guerin F (2019) End-to-end sequential metaphor identification inspired by linguistic theories. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 3888–3898

Miltsakaki E (2007) A rethink of the relationship between salience and anaphora resolution. In: Proceedings of the 6th discourse anaphora and anaphor resolution colloquium, pp 91–96

Mitamura T, Liu Z, Hovy EH (2016) Overview of TAC-KBP 2016 event nugget track. In: Proceedings of the 2016 text analysis conference, TAC 2016, Gaithersburg, November 14–15, 2016. NIST

Mitamura T, Liu Z, Hovy EH (2017) Events detection, coreference and sequencing: what's next? Overview of the TAC KBP 2017 event track. In: TAC

Mitkov R (1999) Anaphora resolution: the state of the art. Citeseer

Moosavi NS, Strube M (2016) 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, pp 632–642. Association for Computational Linguistics

Murugesan K, Atzeni M, Kapanipathi P, Shukla P, Kumaravel S, Tesauro G, Talamadupula K, Sachan M, Campbell M (2021) Text-based RL agents with commonsense knowledge: new challenges, environments and baselines. In: Thirty fifth AAAI conference on artificial intelligence

Ng V (2010) Supervised noun phrase coreference research: The first fifteen years. In: Proceedings of the 48th annual meeting of the association for computational linguistics, Uppsala, pp 1396–1411. Association for Computational Linguistics

Oberle B (2018) SACR: a drag-and-drop based tool for coreference annotation. In: NCC Chair, Choukri K, Cieri C, Declerck T, Goggi S, Hasida K, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, PiperidisS, Tokunaga T (eds.), Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), Miyazaki. European Language Resources Association (ELRA)

O'Gorman T, Wright-Bettner K, Palmer M (2016) Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In: Proceedings of the 2nd workshop on computing news storylines (CNS 2016), Austin, pp 47–56. Association for Computational Linguistics

OpenAI (2022) Introducing ChatGPT

OpenAI (2023) GPT-4 technical report

Peng H, Chang KW, Roth D (2015) A joint framework for coreference resolution and mention head detection. In: Proceedings of the nineteenth conference on computational natural language learning, Beijing, pp 12–21. Association for Computational Linguistics

Pennington J, Socher R, Manning C (2014) GloVe: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, pp 1532–1543. Association for Computational Linguistics

Poesio M, Stuckardt R, Versley Y (2016) Anaphora resolution-algorithms, resources, and applications. Theory and applications of natural language processing. Springer, New York

Pradhan S, Moschitti A, Xue N, Uryupina O, Zhang Y (2012) CoNLL-2012 shared task: modeling multilingual unrestricted coreference in OntoNotes. In: Joint conference on EMNLP and CoNLL - Shared Task, Jeju Island, pp 1–40. Association for Computational Linguistics

Raghunathan K, Lee H, Rangarajan S, Chambers N, Surdeanu M, Jurafsky D, Manning C (2010a) A multipass sieve for coreference resolution. In: Proceedings of the 2010 conference on empirical methods in natural language processing. Cambridge, MA, pp 492–501. Association for Computational Linguistics

Raghunathan K, Lee H, Rangarajan S, Chambers N, Surdeanu M, Jurafsky D, Manning C (2010b) A multipass sieve for coreference resolution. In: Empirical methods in natural language processing (EMNLP)

Rahman A, Ng V (2012) Resolving complex cases of definite pronouns: the Winograd schema challenge. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, Jeju, pp 777–789. Association for Computational Linguistics

Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 conference on empirical methods in natural language processing, Austin, pp 2383–2392. Association for Computational Linguistics

Recasens M, Hovy E (2011) BLANC: implementing the rand index for coreference evaluation. Nat Language Eng 17(4):485–510

Reiter N (2018) CorefAnnotator: a new annotation tool for entity references. Data in the Digital Humanities. In: Abstracts of EADH

Riedel S, Yao L, McCallum A, Marlin BM (2013) Relation extraction with matrix factorization and universal schemas. In: Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies, Atlanta, pp 74–84. Association for Computational Linguistics

Ross S, Gordon G, Bagnell D (2011) A reduction of imitation learning and structured prediction to no-regret online learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics, pp 627–635. JMLR Workshop and Conference Proceedings

Rudinger R, Naradowsky J, Leonard B, Van Durme B (2018) 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, pp 8–14. Association for Computational Linguistics

Shlain M, Taub-Tabib H, Sadde S, Goldberg Y (2020) Syntactic search by example. In : Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations, pp 17–23. Association for Computational Linguistics

Stenetorp P, Pyysalo S, Ananiadou S, Tsujii J (2011) Almost total recall: semantic category disambiguation using large lexical resources and approximate string matching. In: Proceedings of the fourth international symposium on languages in biology and medicine. Citeseer

Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J (2012) April. BRAT: a web-based tool for NLP-assisted text annotation. In: Proceedings of the demonstrations at the 13th conference of the european chapter of the association for computational linguistics, Avignon, pp. 102–107. Association for Computational Linguistics

Stoyanov V, Gilbert N, Cardie C, Riloff E (2009), August. 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, Suntec, Singapore, pp 656–664. Association for Computational Linguistics

Sukthanker R, Poria S, Cambria E, Thirunavukarasu R (2020) Anaphora and coreference resolution: a review. Inform Fus 59:139–162

Sun Y, Wang S, Feng S, Ding S, Pang S, Shang J, Liu J, Chen X, Zhao Y, Lu Y, Liu W, Wu Z, Gong W, Liang J, Shang Z, Sun P, Liu W, Ouyang X, Yu D, Tian H, Wu H, Wang H (2021) ERNIE 3.0: large-scale knowledge enhanced pre-training for language understanding and generation. CoRR abs/2107.02137. arXiv:2107.02137

Teh Y, Bapst V, Czarnecki WM, Quan J, Kirkpatrick J, Hadsell R, Heess N, Pascanu R (2017) Distral: robust multitask reinforcement learning. In: Advances in neural information processing systems, pp. 4496–4506

Thirukovalluru R, Monath N, Shridhar K, Zaheer M, Sachan M, McCallum A (2021) Scaling within document coreference to long texts. In: Findings of the association for computational linguistics: ACL-IJCNLP 2021, pp 3921–3931. Association for Computational Linguistics

Turian J, Ratinov LA, Bengio Y (2010), July. Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th annual meeting of the association for computational linguistics, Uppsala, pp 384–394. Association for Computational Linguistics

Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR (2012) Evaluating the state of the art in coreference resolution for electronic medical records. JAMIA 19(5):786–791

Varkel Y, Globerson A (2020) Pre-training mention representations in coreference models. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 8534–8540. Association for Computational Linguistics

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008

Verga P , McCallum A (2016) Row-less universal schema. In: Proceedings of the 5th workshop on automated knowledge base construction, San Diego, pp 63–68. Association for Computational Linguistics

Vilain M, Burger J, Aberdeen J, Connolly D, Hirschman L (1995) A model-theoretic coreference scoring scheme. In: Sixth message understanding conference (MUC-6): proceedings of a conference held in Columbia, Maryland, November 6–8, 1995

Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O, Bowman S (2019). SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc, F, Fox E, Garnett R (eds.), Advances in neural information processing systems, volume 32. Curran Associates, Inc

Wang Y, Shen Y, Jin H (2021) An end-to-end actor-critic-based neural coreference resolution system. In: ICASSP 2021-2021 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 7848–7852

Webster K, Recasens M, Axelrod V, Baldridge J (2018) Mind the GAP: a balanced corpus of gendered ambiguous pronouns. Trans Assoc Comput Linguist 6:605–617. https://doi.org/10.1162/tacl_a_00240

Welbl J, Stenetorp P, Riedel S (2018) Constructing datasets for multi-hop reading comprehension across documents. Trans Assoc Comput Linguist 6:287–302

Winograd T (1972) Understanding natural language. Cognit Psychol 3(1):1–191. https://doi.org/10.1016/0010-0285(72)90002-3

Wiseman S, Rush AM, Shieber SM (2016) Learning global features for coreference resolution. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, San Diego, pp 994–1004. Association for Computational Linguistics

Wu W, Wang F, Yuan A, Wu F, Li J (2020) CorefQA: coreference resolution as query-based span prediction. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 6953–6963. Association for Computational Linguistics

Xia P, Sedoc J, Van Durme B (2020) Incremental neural coreference resolution in constant memory. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 8617–8624. Association for Computational Linguistics

Xu L, Choi JD (2020) Revealing the myth of higher-order inference in coreference resolution. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 8527–8533. Association for Computational Linguistics

Yadav N, Monath N, Angell R, McCallum A (2021) Event and entity coreference using trees to encode uncertainty in joint decisions. In: Proceedings of the fourth workshop on computational models of reference, anaphora and coreference, Punta Cana, pp 100–110. Association for Computational Linguistics

Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) XLNet: generalized autoregressive pretraining for language understanding. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds.), Advances in neural information processing systems, volume 32. Curran Associates, Inc

Ye D, Lin Y, Du J, Liu Z, Li P, Sun M, Liu Z (2020) Coreferential reasoning learning for language representation. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 7170–7186. Association for Computational Linguistics

Yu J, Bohnet B, Poesio M (2020) Neural mention detection. In: LREC

Yu X, Yin W, Roth D (2020) Pairwise representation learning for event coreference

Zeldes A (2017) The GUM corpus: creating multilayer resources in the classroom. Lang Resour Eval 59:581–612. https://doi.org/10.1007/s10579-016-9343-x

Zeng Y, Jin X, Guan S, Guo J, Cheng X (2020) Event coreference resolution with their paraphrases and argument-aware embeddings. In: Proceedings of the 28th international conference on computational linguistics, Barcelona, pp 3084–3094. International Committee on Computational Linguistics

Zhang H, Song Y, Song Y, Yu D (2019) Knowledge-aware pronoun coreference resolution. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, pp 867–876. Association for Computational Linguistics

Zhang R, Nogueira dos Santos C, Yasunaga M, Xiang B, Radev D (2018) 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, pp 102–107. Association for Computational Linguistics

Zhang X, Zhao J, LeCun Y (2015) Character-level convolutional networks for text classification. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R (eds) Advances in neural information processing systems, vol 28. Curran Associates Inc, Red Hook

Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW (2018) 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, pp 15–20. Association for Computational Linguistics

Zhu P, Zhang Z, Li J, Huang Y, Zhao H (2018) Lingke: a fine-grained multi-turn chatbot for customer service. In: Proceedings of the 27th international conference on computational linguistics: system demonstrations, Santa Fe, pp 108–112. Association for Computational Linguistics

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.