Artificial Intelligence Review (2021) 54:2917–3006
https://doi.org/10.1007/s10462-020-09917-3

A comprehensive review on feature set used for anaphora resolution

Kusum Lata[1] · Pardeep Singh[1] · Kamlesh Dutta[1]

Published online: 14 October 2020

Abstract

In linguistics, the Anaphora Resolution (AR) is the method of identifying the antecedent for anaphora. In simple terms, this is the problem that helps to solve what the expression referring to a referent refers to. It is considered to be one of the tedious tasks in Natural Language Processing (NLP). AR's burgeoning popularity among researchers is attributable to its strong relevance to machine translation, text summarization, chatbot, question answering, and many others. This paper presents a review of AR approaches based on significant features utilized to perform this task and presents the evaluation metrics for this field. The feature is a relevant term related to AR that provides vital information regarding anaphor, antecedent, and relation between them. In this context, features represent the lexical, syntactical, semantical, and positional relationship between anaphor and its possible candidate antecedent. The performance of the Anaphora resolution system is profoundly dependent on the features used in the AR system. Hence, the selection of features for the AR system is highly significant. The main emphasis is to provide an overview of the various features needed to extract both the Anaphora and the Antecedent, respectively, used in different AR systems, present in literature. It is observed that syntactical information enhances the correctness of determining the properties for the existence of an anaphor and antecedent identification. Nowadays the trend is changing from hand-crafted feature dependent methods to deep learning approaches which try to learn feature representation. The performance of deep learning is progressing due to the accessibility of additional data and more powerful computing resources. This survey will provide the state-of art for the better understanding of solving AR problem from the feature selection perspective. The findings of this survey are useful to provide valuable insight into present trends and are helpful for researchers who are looking for developing AR system within given constraints.

Keywords Anaphora · Anaphora resolution · Anaphor · Antecedent · Feature set · Feature selection · Natural language processing

✉ Kusum Lata
ranapoo@gmail.comAffiliations

Pardeep Singh
pardeep@nith.ac.in

Kamlesh Dutta
kd@nith.ac.in

[1] Computer Science & Engineering Department, National Institute of Technology Hamirpur, Hamirpur, Himachal Pradesh, India

1 Introduction

Humans use languages to express and convey their thought in many forms and contexts in daily life. Human beings must be coherent in passing their thoughts on to another human being successfully. Such coherent constructs are generally called Discourse, and can be represented by a series of sentences. Coherence, when explaining events as they happened, is not always present in a linear manner. Consequently, a construct may be coherent even though it does not obey the sequence of events that have taken place. To make these constructs coherent, there is a need to make sure that they are cohesive, i.e., how they are linked and makes a meaningful sense. Humans communicate text in an unstructured way. Therefore, much of the world's information is available in unstructured form, i.e., raw text in English or some other language. A human can interpret the meaning of the written and spoken text with their intelligence. Still, the machine/computer cannot easily interpret the meaning of the text, i.e., human language. The area of research that focuses on human language-computer interaction called NLP is a subfield of Artificial Intelligence (AI), which is concerned with the interactions between computers and human languages. Anaphora Resolution (AR) is one of the most inspiring research areas in Natural language processing. Anaphora often occurs in written and spoken text. It is evident that text is indeed not only the strings of the sentence; there could be multiple statements with many nouns and pronouns with their different references. AR deals with the process of finding the antecedent for anaphora. Some of the terms are explained as below:

Antecedent: Something previous or pre-existing.
Anaphora: This is made up of two words viz,
ANA meaning back or upstream and
PHORA meaning Act of Carrying. Therefore
ANAPHORA means "Act of Carrying Backward"
Anaphora Resolution: To find the antecedent of given anaphora that refers to the same entity.

For example

E1: *"Chief Minister Vijay Rupani has announced that he would be camping in the district for next five days to oversee relief and rehabilitation work."*
[Times of India, Jul 31, 2017]

In the above example, "*he*" is the anaphor which is referring to the antecedent "*Chief Minister Vijay Rupani*".

Interpreting anaphora is straightforward; however, the text also contains multiple antecedents, which should be resolved for the correct anaphor. Anaphors with antecedents are known as anaphoric, and those without are not anaphoric. AR is a challenging task in the NLP, because it needs not only knowledge but also competence in all areas of language processing, such as lexical, morphological, syntactic, semantic, etc. Pragmatic knowledge is often essential in AR, as discussed by the authors (Carbonell and Brown 1988).

As reported by the form of anaphora, there are many varieties of anaphora (Mitkov 2014): *Lexical noun phrase* (NP) anaphora perceived syntactically as a *definite noun* *phrase*, *pronominal anaphora* perceived by Anaphoric pronouns, *Adjectival anaphora* perceived by anaphoric possessive adjectives. Anaphora can be categorized according to the antecedent location, which is intra-sentential or inter-sentential:

- *Intra-Sentential:* says the anaphor and antecedent belong to the same sentence
- *Inter-Sentential:* says the anaphor and antecedent does not belong to the same sentence

1.1 Anaphora resolution framework

In general, AR system consists of the following modules namely, *Features Identification*, *Feature Extractor module, Feature Selection*, and *AR module*. Figure 1 shows the generic anaphora resolution framework, adopted for anaphora resolution developed in different languages. To resolve anaphora, the identification of anaphora and antecedent is required.

- *The Feature Identification Module*: This module finds features to identify both the anaphora and the antecedent. Further, it also finds features to identify the relation between anaphora and antecedents.
- *Feature Extractor module:* This Module extracts the features identified by the identification module by tokenization, sentence splitter, and morphological analyzer, and part of speech tagger, semantic/syntactic parser, named entity recognizer. All the features are then available for use after this module.
- *Feature Selection:* This is then used to select features from a large set of features. For example, we have 'N' number of features, and after using the feature selection method, it can be reduced to 'M' number of features.
- *Anaphora Resolution module:* This module then uses selected features to find the correct antecedent of anaphor.

1.2 Feature set for anaphora resolution

Anaphora Resolution approaches are largely dependent on the features used in the resolution process, ranging from syntactic features, semantic features, and other deep knowledge. Identification of feature set is one of the critical steps in resolving anaphora. A suitably annotated corpus, especially focusing on anaphora is the first and foremost prerequisite to obtain desired feature set for the purpose of anaphora resolution. Large numbers of features reduce the generalization capabilities and add redundant and probably superfluous details, while a limited set of features may be inadequate for resolution. However, regardless of the approach to anaphora resolution, some features are important while others may only be significant. Several researchers have tried to reduce the multidimensionality of the feature set for Anaphora Resolution Algorithm to address computational cost (Arregi et al. 2010a, b; Shekhar et al. 2018). Numbers and categories of features used determine the complexity of the problem. One could argue that more information would produce a better result. However, looking at the relevance of problem of anaphora resolution, use of more features may only increase the computational time and may not make much of the difference in terms of accuracy. The results may be good in some situations, but within a critical limit, the use of new features may result in unintended problems, and in all probability errors may occur. The problem of redundancy due to the similarity between features affects all anaphora resolution procedures equally, resulting in increased processing times as well as uncertainty of inter classes caused by the introduction of excessive or misleading features. Performance of any anaphoric resolver greatly depends on the quality of a high accurate mention detector and the use of appropriate features for anaphora resolution. The choice and selection of feature also depend on the richness of the language in terms of NLP. The languages having extensive research work available have the luxury of utilizing large number of features, whereas resource-poor languages tend to work with ‘low-cost’ learning features readily obtainable from the output of a part-of-speech tagger, and largely bypass deep syntactic and semantic analysis (Cuevas and Paraboni 2008). In recent years, a general change from creating handcrafted feature set to automated generation is prevalent due to the availability of machine learning approaches. Based on these observations it becomes pertinent to focus our research on the ways features have been identified, extracted and selected in various anaphora resolution approaches. Also, it’s important to understand the ways by which feature sets are generated in different languages. Many a times the choice of feature set gets influenced by the main NLP application, where the anaphora resolution is to be included as a wrapper or a plug-in. This paper aims to find the answer to five Research Questions (RQs) with respect to the feature set used by various researchers in anaphora resolution approaches developed for various languages:

(1) RQ1: What are the most relevant/sufficient/essential features for AR?
RQ1 aims at identifying the essential features of AR approaches.
(2) RQ2: How to extract and select efficient features adapting to different approaches used for AR?
RQ2 aims for extracting and selecting features used differently for anaphora resolution approaches.
(3) RQ3: Is there any correlation among features used in the AR system?
RQ3 is to find the association between features used in the AR system.
(4) RQ4: What is the relationship between features and the AR approach used?
RQ4 covers the relationship between features and AR approaches.
(5) RQ5: Which features are languages specific that are used in the AR system?
RQ5 aims to identify the language-specific features.

1.3 Contribution of the paper

The research work reported in this paper shall be useful for NLP researchers in making choices for the feature set in respect of an anaphora resolution problem. The key contributions of the paper are as follows:

- 274 papers have been reviewed in this survey, and a detailed summary of features selected for various Anaphora Resolution Algorithms developed for various languages is presented.
- We highlight various NLP applications along with real-world applications of anaphora resolution.
- We present previous surveys for anaphora resolution, carried out by various researchers.
- Forty-one datasets used in the AR task are discussed from the feature selection point of view.
- We provide an empirical analysis of the performance of selected models, highlight the improvements in the performance of the approaches, and sets the state-of-the-art on AR to this date.

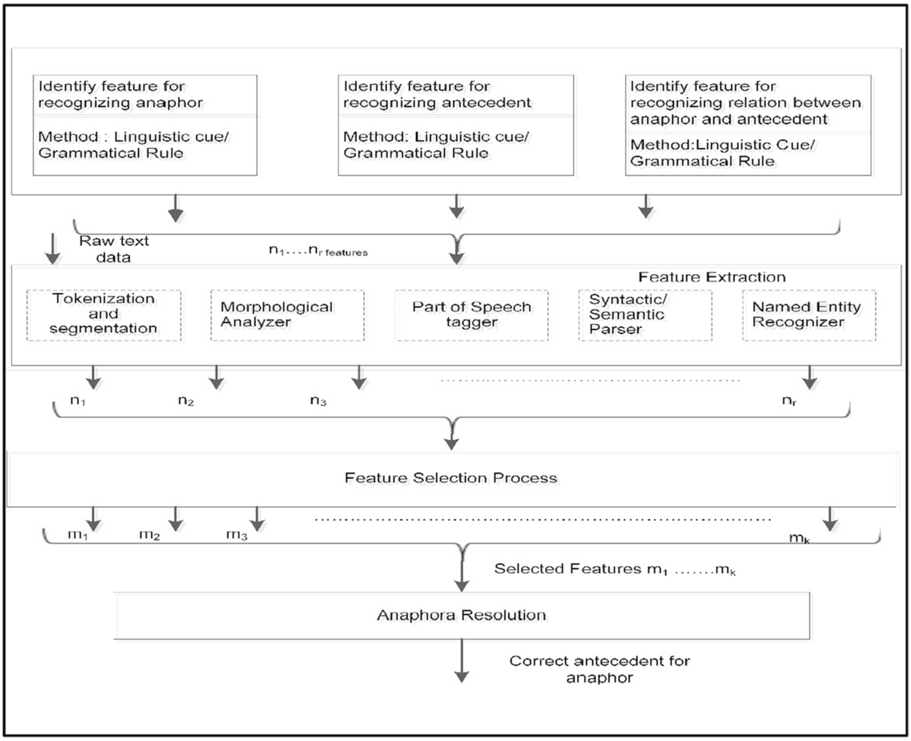
Identify feature for recognizing anaphor
Method : Linguistic cue/ Grammatical Rule
Identify feature for recognizing antecedent
Method: Linguistic cue/ Grammatical Rule
Identify feature for recognizing relation between anaphor and antecedent
Method:Linguistic Cue/ Grammatical Rule
Raw text data
n1....nr features
Feature Extraction
Tokenization and segmentation
Morphological Analyzer
Part of Speech tagger
Syntactic/ Semantic Parser
Named Entity Recognizer
n1
n2
n3
nr
Feature Selection Process
m1
m2
m3
mk
Selected Features m1 ....mk
Anaphora Resolution
Correct antecedent for anaphor

Fig. 1 Anaphora resolution system

- We discuss challenges and future scope in the field of AR.

The remaining paper is structured under the following sections: Sect. 2 highlights various NLP applications reported in the literature that have focused on the inclusion of anaphora resolution and its overall impact on the accuracy and performance of the respective NLP application. Section 3 covers previous surveys carried out on AR. Section 4 provide the details of the dataset available for different languages such as English, Hindi, German, Czech, etc. Section 5 details the evaluation metrics used for evaluating the AR algorithm. Section 6 describes the different features for AR task being used in detail. Section 7 provides information about various methods of feature selection. In Sect. 8, the selection of different features for distinct AR approaches are discussed. Section 9 includes the discussion about the performance of AR tasks, including Gendered pronoun resolution. Section 10 summarize the limitation of AR approaches also highlights the answer to research questions (RQ). In this section, the need of AR in Text summarization, one of the NLP application, is also highlighted. Section 11 discusses the Challenges and Future Scope. The conclusion of this paper is summarized in Sect. 12.

2 Anaphora resolution in NLP applications

Anaphora Resolution is relevant in several NLP tasks such as Information Extraction (IE), Machine Translation (MT), Question Answering (QA), Sentiment Analysis, and Text Summarization. Significant research in the area of anaphora resolution has primarily focused on the development of various approaches to resolve anaphor in the text, ranging from traditional approaches, to machine learning (including deep learning). Some NLP applications have attempted to demonstrate how anaphora resolution is capable of improving the overall performance of that application. In an ideal situation, anaphora resolution, if done without introducing errors, always improve the performance of any NLP application. Nevertheless, if resolution of anaphora could not be achieved with sufficient precision, it could pose far more problems than it could solve. Most researchers have found that although the resolution of anaphora is important for the NLP application, its impact is limited. Researchers have advised not to introduce anaphora resolution in the NLP application unless it is essential. In this section we briefly describe various NLP applications, reported in literature where anaphora resolution has been included with varied success.

- *Information Extraction* is the process of extracting structured information automatically from machine-readable unstructured and/or semi-structured documents such as entity detection, relation extraction between entities, etc. The use of anaphora resolution in the information extraction system is explored in these papers: (Wang et al. 2002; Hahn et al. 2002; Kobayashi et al. 2005; Mur and Van Der Plas 2006; Matysiak 2007). The application of AR in relation extraction that is a subtask of IE has also been explored (Kilicoglu et al. 2016) in which they demonstrated that a heavily semantic approach to sortal anaphora resolution is highly effective for biomedical literature. The utilization of AR information in the recommendation system is explored (Wohiduzzaman and Ismail 2018). They developed a recommendation system for the Bangla News domain with the help of anaphora resolution to increment the keywords frequency. Ting et al. (2019) also exploited the anaphoric information in named entity recognition that is a subtask of IE and showed that the performance of identifying entities for the person class.
- *Machine Translation:* MT refers to the automated translation. This is how computer software is used to translate a text or speech from one natural language to another. MT performs mechanical replacement of words from one language for words in the other, however, that by itself once in a while create a decent interpretation since acknowledgment of entire expressions and their nearest partners in the objective language is required. The issues on AR were discussed (Mitkov 1999a) in Machine Translation and Multilingual NLP. The understanding of anaphora is fundamental for Machine Translation System to perform successfully. It is crucial to resolve the anaphoric relation while the translation into the languages which mark the gender of pronouns. The application of AR in machine translation has also been explored in these papers: (Nakaiwa and Ikehara 1992; Mitkov 1995; Mitkov and Schmidt 1998; Mitamura et al. 2002; Dutta et al. 2009; Hardmeier and Federico 2010; Novák 2011; Loáiciga and Wehrli 2015; Voita et al. 2018; Stojanovski and Fraser 2019).
- *Question Answering:* This is another very extensively used application of NLP. This is involved in building systems that automatically answer questions posed by humans in a natural language. The Stanford Question Answering Dataset (SQuAD), which is a new reading comprehension dataset comprises of questions on a set of Wikipedia articles

described by Rajpurkar et al. (2016) and also included the concept of AR. The traditional work also proffered meaningful results as per that era. The effect was analyzed (Vicedo and Ferrández 2000a) using pronominal AR to Question Answering (QA) systems and showed that QA performance improves by applying it. The following papers provide a superior explanation for the application of AR in QA: (Vicedo and Ferrández 2000b; Castagnola 2002; Watson et al. 2003). (Zhao et al. 2019b) presents a good work for NLP applications, in which they evaluated the capsule network approach on two tasks: Multi-label Text Classification and Question Answering. Capsule networks helped them achieving remarkable results in both the fields in low resource settings with few training instances.

- *Sentiment Analysis:* This field helps in finding the sentiments of a user from a text like a review, comment, and more. This task describes the emotional connection behind a particular text, such as positive, negative, or neutral. Ma et al. (2018a) developed an attentive LSTM which is the extension of Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) for Targeted aspect-based sentiment analysis via injecting the commonsense knowledge. Ma et al. (2018b) developed the hybrid neural network that injects semantic information into deep neural networks for Targeted Aspect-Based Sentiment Analysis. They suggested the extension of LSTM to incorporate the explicit knowledge with implicit knowledge and this extended model is called as Sentic LSTM and additionally, also proposed hybrid of Sentic LSTM and recurrent additive neural network (H-Sentic-LSTM) to exploit the concept level inputs. Sentiment can be classified based on the aspect as described by Wang et al. (2016) and Zeng et al. (2019). Zeng et al. (2019) utilized attention-based LSTM for aspect-level sentiment classification to exploit the information of the explicit position context. Do et al. (2019) provided an overview of the major deep learning approaches and a precise comparison of these approaches for sentiment analysis at the aspect level. The novel trends and methods using deep learning approaches (Habimana et al. 2020) were presented, in which the study was on the recent trends in Sentiment Analysis using Deep Learning approaches and their variants too, and then applied them onto the different sentiment analysis tasks. The study Opinion Mining using anaphora resolution (Jakob and Gurevych 2010) was to find the motive behind users' movie reviews. The methodology used was the rule-based CogNIAC AR algorithm. The findings showed favorable results regarding the extraction of opinion targets. A deep briefing about how AR can be used in the real world towards Sentiment Analysis (Nithya 2019) was presented. It describes that once the resolution is made, it can predominantly advance the accuracy of supporting sentiment analysis. The importance of determining the shell noun is conceived and expressed using Family-Shell mapping.
- *Text Summarization:* Text summarization refers to the method of compressing long pieces of text to short. The purpose is to create a coherent and fluent summary that includes only the important and main points that are outlined in the document. Automatic Cohesive Summarization with Pronominal Anaphora Resolution (Antunes et al. 2018) was in which a new method called Anaphoric Expression Solver (AES) was presented for extractive text summarization that endeavors to generate more coherent summaries by unraveling pronominal anaphoric expressions. The AES method was able to achieve correct coreference substitution improving the cohesion of equivalent to 81% of the total evaluated. Steinberger et al. (2007) utilized the lexical and automatically extracted anaphoric information in LSA based text summarization. The works in this area are explored in these papers: (Kabadjov et al. 2005; Orasan and St 2007; Steinberger et al. 2007; Rusu et al. 2009; Batista et al. 2018; Kameyama 1997) to develop

robust and an efficient reference resolution algorithm for an end-to-end state-of-the-art text summarization system.

The recent progress in deep learning approaches for text classification tasks viz. Sentiment Analysis, Question Answering, News Categorization, and Natural Language Inference is described by Minaee et al. (2020). The authors also analyzed the different deep learning model on the benchmark datasets of different NLP tasks to provide an extremely scalable and secure framework. Though the above-mentioned applications are well discussed by researchers as an NLP task, the real-world application of anaphora processing demands for much more attention and requires the inclusion of domain-related information (Stuckardt 2016) as well. Recognizing that the traditional approaches built in the preceding decades are focused on deep domain-specific knowledge and therefore lack the degree of scalability and robustness, demanded by real-world applications such as biomedicine, medicine, and legal domain, the new trend is to go for a detailed understanding of the particular application by including additional resources such as corpora annotated with domain information and domain entity databases for case-specific anaphora processing. Some of the notable successful application domains are Biomedicine (Castano et al. 2002; Cooper and Kershenbaum 2005; Eilbeck et al. 2005; Gasperin 2006; Vlachos and Gasperin 2006; Vlachos et al. 2006; Gasperin and Briscoe 2008, Cohen et al. 2010), flyBase curator (Karamanis et al. 2008), clinical domain (Savova et al. 2011; Zheng et al. 2011; Wang et al. 2011; He 2007), legal (Dozier and Zielund 2004; Gupta et al. 2018; Dariescu and Gîfu 2019), spatial domain (Leidner 2008; Leidner et al. 2003), news (Popescu-Belis and Lalanne 2004; Kuo and Chen 2004; Stede et al. 2006; Hong and Park 2004) and security (Hu et al. 2020).

3 Previous surveys

A term called Coreference Resolution (CR) is also of the utmost importance when talking about AR. CR in linguistics is a task of finding out all the expressions that refer to the same entity. The scope of the study is limited to AR because nominal anaphora includes various phenomenon such as coreference (Hobbs 1979), bridging (Clark 1975), and comparative anaphora (Halliday and Hasan 1976).

In the last three-decades, AR has been widely discussed and prominently published in English and on Indian languages. A few of them have been referred to in this section for a thorough analysis of AR.

In English, (Mitkov 1999b) presented the first significant exhaustive survey for the AR system only. The author in this paper presented a comprehensive description of the available types of anaphora in text, essential factors considered during the AR process, and various major AR algorithms. The author also focused on the utilization of AR in other related fields within NLP.

Consequently, a comprehensive review of the computational linguistics model of AR (Poesio et al. 2011) was presented and covered the work done in the past forty years. The author also reviewed the CR approaches such as data-driven, machine learning method, annotated corpora containing anaphoric information, metrics used to evaluate the approaches used to solve AR and CR task.

Kolhatkar et al. (2018) provided an exhaustive survey for different approaches used for non-nominal-antecedent (non-NA) anaphora resolution. They categorized the approaches into two categories: Rule-based approach and Machine learning approach (including deep learning techniques). The authors also defined the phenomenon of non-NA anaphora from syntactic and semantic perspectives, description of the linguistic properties of non-NA anaphora, described the annotation efforts and major resources concerning different aspects.

The next survey (Sukthanker et al. 2020) addressed the progress of research in the related field of AR and coreference resolution (CR) and analyzed the resolution algorithm on different datasets. The authors categorized approaches briefly based on techniques used in AR and CR systems. The categorization for entity resolution (AR and CR) was done by the authors in the succeeding three categories viz. Rule-based approach (Hobbs 1978; Carter 1986; Lappin and Leass 1994; Liang and Wu 2004), Statistical based approach (Aone and Bennett 1995; Ge et al. 1998; Mitkov et al. 2002; Mitkov 2001), Deep Learning-based approach. The authors also explained the distinction between CR and AR task and mainly analyzed the algorithm which used recent deep learning techniques for CR task.

Another survey for entity coreference resolution (Stylianou and Vlahavas 2019) effectively addressed the gaps by presenting a detailed review of the neural entity coreference resolution. The research also presents a summary of gendered pronoun resolution that is a subtask of coreference resolution.

The recent novel survey (Kocijan et al. 2020) was performed, presenting a review of the existing Winograd Schema Challenge benchmark dataset and explaining the different kinds of approaches that are implemented to solve the Winograd Schema Challenge (resolve ambiguous pronoun). The authors categorized approaches into three categories: *Feature-based approaches, Neural network approaches, and Language Model approaches*. Feature-based approaches utilized a set of rules created from knowledge bases, internet search queries and used the logic-based system to solve the task; Neural approaches utilized neural network architecture and deep learning techniques excluding the use of pre-trained language models. Language Model approaches utilized the language model based on neural networks to address the Winograd Schema Challenge.

In the Indian language, there is a review focused on AR for Indian languages done by the author (Yadav et al. 2016). The authors categorized the AR approaches into four categories: Rule-based, Corpus-based, Knowledge-poor, Discourse based approaches. In Hindi language, Mahato et al. (2019) presented a survey on AR approaches and several features used for these approaches. The author also reviewed the computational behavior of AR approaches.

All the above listed surveys can be taken as a reference to understand the erstwhile approaches in AR and CR. The previous surveys mostly emphasized on a single language i.e., English for AR and CR task, and few researchers focused on survey for Indian languages such as the Hindi language. Also, these surveys have not focused on the work related to the feature set used, its multidimensionally character and variation resulting due to the different languages and the particular NLP task where anaphora resolution has been employed. To our knowledge, a systematic survey of AR is yet to be done, which consider the relevance of feature set used in different NLP applications and also different languages.

4 Datasets for anaphora resolution

This section presents the dataset that contains annotation for the relation between the anaphors and their antecedents. The datasets were developed for the task of anaphora resolution in years for various languages such as English, Chinese, Arabic, German, Catalan, Dutch, French, Italian, Japanese, and Spanish, etc. Based on the past researches, it was observed that the dataset mostly being used for AR task, was based on different features such as the domain, the annotation schemes, and the types of references that are labeled, etc. Therefore, it becomes essential to have a clear understanding of the AR datasets before continuing with the research strategies relating to the tasks which would be utilizing these datasets either for training or for inferring effective rules. Various AR datasets used for the AR task are presented along with the problem handling scenario and classified according to the text-domain.

4.1 News classification datasets

Brown corpus[1] (Kučera and Francis 1967) was created at Brown University and was the first text corpus of American English consisting of 1 million words of 15 different text categories such as News Texts on different topics, non-fiction and fiction books along with government documents. This corpus although small, is still used.

The SUSANNE Corpus[2] (Sampson 2002) is an annotated corpus which consists of 130,000-word cross-section of written American English and subset of Brown Corpus. It is freely available for researchers.

MUC[3] (MUC-3 (Sundheim 1991)/MUC-4(Sundheim 1992)) dataset is mostly used for comparison of an anaphora resolution algorithm. MUC-3 and MUC-4 corpus consist of Latin American terrorism news articles. This was the first corpus created for the AR task for the English language. Anaphoric Treebank corpus consists of subsamples of the Associated Press Treebank (AP) corpus (Leech and Garside 1991), which contains pronominal anaphoric annotation with UCREL anaphora annotation scheme and also show reference to lexical cohesion and consist of 100,000 words. AP corpus consists of American newswire reports.

The Korean Treebank (Han et al. 2001) is an annotated corpus at the level of morphological and syntactic features and comprises of military language training manuals texts containing 54,366 words and 5078 sentences. Korean Treebank Version 2.0 is an extension of the Korean English Treebank Annotations corpus; Korean Treebank 2.0 consists of Korean news text.

Prague Dependency Treebank2.0[4]comprises text from three sources viz, Newswire, News Magazine, Journal article text for the Czech language, which was annotated at different levels of morphological, syntactic, and complex semantic annotation. Specific properties of sentence information structure and coreference relations are annotated at the semantic level. Prague Dependency Treebank (Hajic 1998) is the result of various evaluation series from 2001 to 2018 and contains a large amount of text for the Czech language from a different genre: Newswire, News Magazine, Journal. It is manually annotated at the level of Morphological, Analytical, Techtogrammatical. Techtogrammatical annotation level contains an anaphoric link.

[1] http://korpus.uib.no/icame/manuals/BROWN/INDEX.HTM.

[2] https://www.grsampson.net/SueDoc.html.

[3] https://www-nlpir.nist.gov/related_projects/muc/.

[4] https://catalog.ldc.upenn.edu/LDC2006T01.

The TuBa-D/Z Treebank (Hinrichs et al. 2005) is syntactically annotated manually, or rather semi-automatically which contains German news text. This treebank also contains coreference and anaphoric annotation. The 11.0 version of this corpus is also available in CoNLL-U(v2) format.

AnCora-CO[5] (Recasens and Martí 2010) comprises of a corpus for Catalan (AnCora-CA) and a corpus for Spanish language (AnCora-ES).Each of these is made up of 500 k words and contains a coreference link between pronouns, full NP, and discourse segments annotated news text. The corpus can be used for the AR task. It requires the extraction of the pronouns that are included in an entity for this task. They assume that their antecedent corresponds to the previous mention in the same entity. For both languages, each corpus comprises of news text. AnCora-CO corpus is freely available for use or research purposes.

ICON 2011 dataset[6] is used by authors to evaluate the AR system for Indian languages. This dataset consists of news text.

Kyoto text corpus[7] (Kawahara et al. 2002) contains manually annotated news text with morphological and syntactic annotation for the Japanese language.

NAIST Text Corpus developed by (Iida et al. 2007), which is used for zero-anaphora resolution tasks for the Japanese language, comprising of text from Kyoto text corpus and annotated with co-reference and predicate-argument relations.

4.2 Multi domain datasets

The ACE corpus (Doddington et al. 2004), which started from 2000 to 2008 contains the annotated text for different languages like English, Chinese, and Arabic. Initially, this corpus consisted of news-wire articles but subsequently included broadcast conversations, web-log, UseNet, and conversational telephonic speech too. ACE is mainly annotated for pronominal AR, which follows the TIMEX2 annotation scheme and is evaluated using the ACE-score metric. The ACE corpus is not freely accessible for research purposes and distributed through Linguistic Data Consortium.

The VENEX corpus (Poesio 2004) contains an anaphorically annotated text of the Italian language from news texts and task-oriented dialogues text. The annotation scheme followed by this corpus is the MATE scheme.

Cast3LB corpus (Palomar et al. 2004) is a multilingual corpus 3LB2 comprising of corpora annotated with linguistic information such as morpho-syntactic, semantic, and discourse-pragmatic (anaphora): one for Catalan (Cat3LB), one for Basque (Eus3LB) and one for Spanish (Cast3LB).

OntoNotes corpus 5.0 (Hovy et al. 2006) is the largest existing manual annotated corpus with coreference annotations. This corpus comprises of gold POS labels and syntactic constituent parses, along with coreference resolution for pronominal anaphora, definite or proper noun, and named entity annotations for proper nouns. This corpus consists of annotated text from telephonic conversations, newswire, newsgroups, broadcast news, broadcast conversation, weblogs for English, Mandarin Chinese, Arabic, Chinese languages. This dataset is freely available through Linguistic Data Consortium.

[5] http://clic.ub.edu/corpus/en.

[6] http://ltrc.iiit.ac.in/icon/2011/nlptools/.

[7] https://github.com/ku-nlp/KyotoCorpus.

Hammami et al. (2009) developed Arabic corpus which consist of news text, technical computers manuals, Tunisian educational book, and a novel. This dataset is annotated by Anaphoric Annotating Tool for Arabic and follows XML-based format.

The Anaphora Resolution and Underspecification (ARRAU) corpus was created by (Poesio and Artstein 2008) for English language. This corpus contains anaphoric annotated text and mainly focusing on plural anaphora, abstract anaphora, ambiguous anaphoric entities. This corpus follows MMAX2 format and contains four several text genres viz. news text from Wall Street Journal dataset (RST),[8] task-oriented dialogues text (TRAINS-91 and TRAINS-93),[9] fiction text (PEAR)[10]and medical leaflets (GNOME) for English language. This dataset is available for research use through Linguistic Data Consortium.

RuCor Corpus (Ju et al. 2014) contains manually anaphorically, coreferential, morphologically annotated text for Russian language from different sources: news, fiction, science, blog, essays, Russian Wikipedia. The annotation scheme is based on MUC. The dataset[11] is freely accessible for research purposes.

Pronoun Disambiguation Problem Dataset[12] (Davis et al. 2017) is used in the first Winograd Schemas Challenge (WCS) and consists of 120 pronoun disambiguation problems that are collected from biographies, autobiographies, popular literature, newspapers (news analysis, news stories). This dataset follows WSC format.

Mahato et al. (2018) developed a dataset for spoken Hindi dialect domain comprising of randomly selected 1757 sentences from different domains such as spoken discourse, including individual statements, conversations between persons, discussions, articles related to everyday subjects from the newspaper.

MuDoCo Corpus (Martin et al. 2020) created a dataset that is composed of authored dialogs between a fictional user and a system that is provided with tasks to perform within six task domains: calling, messaging, reminders, weather, news, and music. The dataset contains a total of 8429 dialogs with an average of 5.36 turns per dialog. The dataset is stored in JSON format and is freely available for research purposes.

4.3 Wikipedia specific domain set

The LIVEMEMORIES corpus (Rodrıguez et al. 2010) adopted the same annotation scheme used by ARRAU corpus. It is a manually annotated corpus of the Italian language for the AR task and consists of the Italian Wikipedia text and blog sites.

WikiCoref corpus (Ghaddar and Langlais 2016) comprises of 30 Wikipedia articles for English language. This dataset mainly emphasizes on a pronominal resolution where candidate antecedent is a nominal mention and follows OntoNotes annotation schema. This corpus is freely available for use.

Georgetown University Multilayer Corpus (GUM) dataset is developed by Zeldes (2017) for English language and consists of approx. 22,500 tokens. It contains a substantial amount of annotated Wikipedia texts of different types such as Interviews, News Articles, Instructional Texts, and Travel Guides and follow CONLL format. It is a freely available

[8] https://catalog.ldc.upenn.edu/LDC2002T07.

[9] https://www.cs.rochester.edu/research/speech/trains.html.

[10] http://pearstories.org/.

[11] http://rucoref.maimbava.net/.

[12] https://cs.nyu.edu/faculty/davise/papers/PDPChallenge.xml.

multilayer corpus of richly annotated web texts. Both Gum and WikiCoref are developed for adapting the Wikipedia data domain.

The new dataset *GAP* (Webster et al. 2018) corpus is created to handle the gender biasness in pronoun resolution and consists of manually annotated Wikipedia text of 8908 ambiguous pronoun–name pairs. This corpus mainly emphasizes on a pronominal resolution where candidate antecedent is a named entity. This dataset is a larger in size and freely available. Thus, this corpus is currently used by Kaggle's Gender Pronoun Resolution competition.[13] It has been observed that some data are mislabeled. The author (Ionita et al. 2019) observed that there is only ~2.5% of instances that were mislabeled.[14]

KNOWREF (Emami et al. 2019) contains the manually annotated text of 8724 Winograd-like text samples that need common sense and word knowledge for pronoun resolution. This corpus comprises of 2018 English Wikipedia, OpenSubtitles, and Reddit comments dating from 2006–2018. This dataset is not freely available and distributed under the CC BY-SA 3.0 license.

New dataset Wikipedia CoREferences Masked (WIKICREM) dataset[15] (Kocijan et al. 2019) is composed of 2.4 M samples of English Wikipedia text (MaskedWiki) to address the issue of insufficiently large training dataset for pronoun disambiguation. The authors also ensure that the named entity appears in not less than twice and masks one of its non-first occurrences. They also make sure the presence of at least one other distinct named entity in the text before the masked occurrence. KNOWREF is freely available for research use.

4.4 Biological specific domain set

MEDSTRACT (Castano et al. 2002) is a larger corpus comprising of MEDLINE abstract only. Two forms of anaphora, namely pronominal and sortal anaphora are primarily concerned in the case of the AR algorithm. This corpus follows the MUC-7 annotation scheme and freely available for research use.

GENIA corpus version 3.0[16] (Kim et al. 2003) comprising of 2000 MEDLINE abstracts containing more than 400,000 words and almost 100,000 annotations for biological terms. It contains biological reactions concerning transcription factors in human blood cells, articles with the MeSH terms, human, blood cell, and transcription factor. This corpus is manually annotated and follows an XML-based mark-up scheme. Also, the corpus is freely available.

The GNOME corpus developed by (Poesio 2004) consists of anaphoric annotated texts from museum labels, pharmaceutical leaflets, and tutorial dialogues with the MATE scheme. The corpus was created to study centering theory. The GNOME corpus is not freely accessible for research purposes.

The FlySlip corpus[17](Gasperin et al. 2007) is a manually annotated corpus for the biomedical domain. It consists of the anaphoric information annotated text of 5 full-text articles that are part of the literature on the molecular biology of Drosophila. This corpus follows a domain-relevant annotation scheme and freely available for researchers.

[13] https://www.kaggle.com/c/gendered-pronoun-resolution.

[14] https://www.kaggle.com/c/gendered-pronoun-resolution/discussion/81331.

[15] https://ora.ox.ac.uk/objects/uuid:c83e94bb-7584-41a1-aef9-85b0e764d9e3.

[16] https://orbit.nlm.nih.gov/browse-repository/dataset/human-annotated/83-genia-corpus.

[17] http://www.wiki.cl.cam.ac.uk/rowiki/NaturalLanguage/FlySlip.

MedCo corpus developed by Su et al. (2008), which comprises of 1999 MEDLINE abstracts from the GENIA data set. This corpus follows the MUC-7 annotation scheme and freely available for researchers.

DrugNerAR corpus[18](Segura-Bedmar et al. 2009) is a corpus created for the biological domain, and the purpose of developing this corpus is to resolve anaphoric expression viz. pronominal, nominal anaphora in drug-drug interactions in pharmacological documents. This corpus consists of manually annotated text collected from the DrugBank annotated database, which contains approximately 4900 entries of drugs. This corpus follows the MUC-7 annotation scheme and freely available for researchers.

The HANNAPIN corpus developed by Batista-Navarro and Ananiadou (2011), which comprises of 20 full-text articles from biochemistry literature and annotated in XML format. In addition to nominal and pronominal anaphora, abbreviations/acronyms as well as numerical coreferences, are usually taken into account. This corpus is freely available for researchers.

4.5 Dialogue specific domain set

The various datasets were developed earlier, although they are not adequate for all the domains because anaphora resolution has its specific characteristics in various domains and text sources. For this reason, dialogue specific domain dataset is created.

SWITCHBOARD corpus (Godfrey et al. 1992) is a telephone speech corpus for research and development and consists of telephone conversations text. This dataset is annotated in the SWBDDAMSL scheme and available for research purposes through LDC with fees.

DARE corpus[19] (Niraula et al. 2014) is created in tutorial dialogue anaphora for the English language. It comprises of anaphorically annotated text collected from conversations between high-school students and the intelligent tutoring system DeepTutor. This data set is manually annotated and freely available to researchers.

4.6 Random text datasets

QurAna corpus (Sharaf and Atwell 2012) developed a large corpus for Arabic Language which consists of Holy Qur'an script. This dataset contains anaphoric link and other anaphoric information. This dataset follows the XML- based format and freely available for researchers through GNU public license.

Winograd Schemas (Levesque et al. 2012) is one of the works related to ambiguous pronoun resolution. This schema consists of approximately 100 examples, which are a pair of sentences. Each pair of sentences has only one or two word differences and has referential ambiguity that resolved with the help of world knowledge and commonsense reasoning information in opposite directions in the two sentences. AI experts create these examples manually. The significant constraint in Winograd schemas was "Google-proof or non-associative as discussed by Trichelair et al. (2018). This challenge is created initially for the English language; it has been translated into other languages viz. French, Portuguese, Japanese, and Chinese and available on the official website[20] of challenge.

[18] http://labda.inf.uc3m.es/DrugDDI/DrugNerAr.html.

[19] http://deeptutor.memphis.edu/resources.htm.

[20] https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html.

Definite Pronoun Resolution corpus[21]*(DPR)* developed by Rahman and Ng (2012), which consists of 943 annotated sentences pair in Winograd schemas, has been written by 30 undergraduate students. The sentences cover a range of subjects viz. actual events, characters, and events in movies, purely imaginary situations. The restriction on the Winograd schemas is eased. Thus, several examples in the dataset are not Google-proof. And then, subsequently, an extension of this data is done by Peng et al. (2015) for general co-reference systems in ACE-2004 format, which is called dataset WinoCoref.[22] They have included the annotation for all pronouns and their linked entities as mentions in this dataset. This dataset did well for obtaining a source of ambiguous pronouns, but it is not generalized and carefully selected.

WinoBias dataset developed by Zhao et al. (2018) consists of 3160 examples which are of pro-stereotype (sentences where a gender pronoun refers to a profession, which is dominated by the same gender) and anti-stereotype subsets (the same set of sentences, but the gender pronoun in each sentence is replaced by the opposite gender).

Hard-CoRe Coreference Corpus developed by Emami et al. (2018) consists of 1275 complicated pronoun disambiguation sentences. This dataset contains short text collected from English Wikipedia and semi-automatic annotated dataset. Specifically, each task instance is a short passage containing a target pronoun that must be correctly resolved to one of two possible antecedents.

WINOGRANDE dataset is a large scale Winograd Schema Challenge created by Sakaguchi et al. (2020), which is consists of 44 k examples collected via crowd sourcing on Amazon Mechanical Turk (AMT) and developed to improve the hardness of problem.

In earlier created datasets viz. ACE dataset, OntoNotes corpus 5.0, etc., there were issues of ambiguous pronoun (more than one candidate antecedent exists). In modern times, there are datasets created for the handling of this limitation in the English language. The dataset as discussed in this Section, Winograd Schemas is the first one created to handle the ambiguous pronoun issues and is a small dataset. The dataset created for handling the ambiguous pronoun resolution are DPR, WikiCoref, WinoCoref, WinoBias, Hard-CoRe Coreference Corpus, WIKICREM, KNOWREF, GAP, WINOGRANDE. GAP is huge in WSC-format than WikiCoref and Winograd Schemas.

5 Evaluation metrics for anaphora resolution task

The performance evaluation of the AR system is an essential component associating with the performance of the algorithm. Initially, the success rate was introduced by (Hobbs 1978) as defined in Eq. (1):

$$Hobbs\ metric(success\ rate) = \frac{Number\ of\ correct\ anaphora\ resolved}{Total\ number\ of\ anaphora\ resolved\ by\ an\ algorithm} \tag{1}$$

Mitkov (2001) defined success rate and critical success rate for evaluating the performance of the AR system which are defined in Eqs. (2), (3) respectively.

[21] http://www.hlt.utdallas.edu/~vince/data/emnlp12/.

[22] https://cogcomp.seas.upenn.edu/page/data/.

$$Success\ rate = \frac{Number\ of\ correctly\ resolved\ anaphors}{Number\ of\ all\ anaphors} \tag{2}$$

$$Critical\ success\ rate = \frac{Number\ of\ correctly\ resolved\ anaphors}{Number\ of\ anaphors\ with\ more\ than\ one\ antecedent\ after\ morphological\ filter} \tag{3}$$

The metrics Precision and Recall can be used to evaluate an AR system. Precision refers to the percentage of relevant results; While Recall, alludes to the percentage of relevant result which has been correctly classified by the algorithm. For evaluating the performance of the system to resolve anaphors, precision and recall can be defined as in Eqs. (4) and (5) respectively. (Baldwin 1997) followed these precisions and recall definitions.

$$Precision(P) = \frac{Number\ of\ correct\ anaphors\ resolved}{Total\ number\ of\ ansphors\ resolved\ by\ algorithm} \tag{4}$$

$$Recall(R) = \frac{Number\ of\ correct\ anaphors\ resolved}{Total\ number\ of\ all\ anaphors\ in\ text} \tag{5}$$

For computing precision and recall, there is a necessity to find out measures called True Positives, True Negatives, False Positives, and False Negatives in the field of AR. These are as follows:

- *True Positive*: indicates an element is correctly recognized as anaphor
- *True Negative*: indicates an element is identified as non- anaphoric that are non-anaphoric
- *False Positive*: indicates an element is identified as anaphor, but it is not anaphoric
- *False Negative*: indicates an element is identified as non-anaphoric, but it is anaphoric

(Aone and Bennett 1995) defined another version of recall computation and defined as the number of correct anaphors resolved divided by the total number of anaphors predicted by the algorithm. F-measure is also introduced as an evaluation metric, which is defined in Eq. (6) where P indicates precision, and R indicates Recall.

$$F\text{−}measure = 2\frac{PR}{P+R} \tag{6}$$

Many evaluation metrics for Coreference resolution task such as (Vilain et al. 1995; Bagga and Baldwin 1998; Doddington et al. 2004; Luo 2005; Recasens and Hovy 2011): *Message Understanding Conferences (MUC),Algorithms for scoring Reference chains* (B3), *ACE metric* (Automatic Content Extraction), *CEAF* (Constrained Entity-Alignment F-Measure), *BLANC* (Bi-Lateral Assessment of Noun-phrase Coreference) exists for evaluating CR system. The detailed explanation of all these metrics (MUC, B3, ACE, CEAF, BLANC) is beyond the scope of this paper since these metrics are well explained and compared in recent literature (Stylianou and Vlahavas 2019; Sukthanker et al. 2020). AR systems are developed for resolving different types of anaphors (Possessive, Reflexive, Third-person pronouns, etc.), and CR task focuses on noun phrase entities. However, there doesn't exist a standard scoring scheme for evaluation of the

AR system because AR algorithms are not evaluated on the same dataset. Few authors (Sobha and Patnaik 2000; Chatterji et al. 2011; Senapati and Garain 2013) used the MUC metric, B3, ACE metric, CEAF, BLANC for evaluation of AR system.

6 Feature set

AR system requires a set of features to resolve anaphora. The feature set is the collection of features representing the properties of individually mentioned entities and entity pairs. It is an essential aspect for selecting informative, distinctive, and independent features for the efficacious algorithm. Mitkov (1999b) suggested that various factors can affect an AR system. Mostly features used are morphological agreement, binding, c-command, semantic similarity, grammatical and semantic parallelism (such as verb similarity, preference, nearness or proximity, etc.). Features were categorized into three classes according to Feature engineering of coreference resolution in multiple languages (Kobdani 2012):

- *Identical features*: these are identical in terms of concept and definition that apply to all languages. String matching is an example of this class.
- *Universal features:* these features have a similar notion to identical features, but a different realization that applies to all languages. The Definiteness of NP, Number can fall in this class.
- *Language-specific features:* these features are limited only to a specific language. For example, Grammatical gender in the German Language.

According to (Dryer 2013), the languages are categorized according to the relative word order as shown in Fig. 2. The word orders can be 'Fix Order' such as English, which has SVO (Subject-Verb-Object) and 'Free order' such as Hindi, Spanish, etc. Further categorization of free word order languages is '*Free Order with One Dominant Order', (*implying one order common in a language like Russian has SVO order as dominant) *and 'Free Order with Lacking Dominant O*rder' (meaning those languages that have no dominant word order such as German, Polish, etc.).

This section covers a summary of the features used in AR. Not all systems need to make use of all these features. Some systems employ a set of substantial features, but some use a small number of features. The following features can be used in many variations:

6.1 Lexical feature

6.1.1 Exact string match

The exact string match feature is used to compare the surface forms of antecedent and anaphor. It means that both anaphor and antecedent should have precisely the same strings sequence. For example,

E2: "*The Q7 is equipped with a 3.0 L supercharged V6 engine which makes 333 hp. Q7's track best is 128 mph.*"
[www.globalcarsbrands.com*, May 31, 2017*]

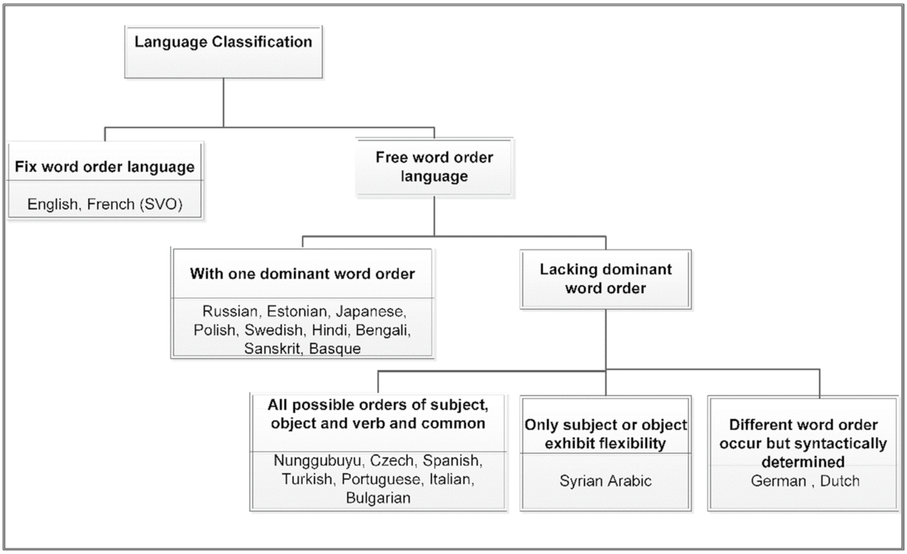
Language Classification
Fix word order language
English, French (SVO)
Free word order language
With one dominant word order
Russian, Estonian, Japanese, Polish, Swedish, Hindi, Bengali, Sanskrit, Basque
Lacking dominant word order
All possible orders of subject, object and verb and common
Nunggubuyu, Czech, Spanish, Turkish, Portuguese, Italian, Bulgarian
Only subject or object exhibit flexibility
Syrian Arabic
Different word order occur but syntactically determined
German , Dutch

Fig. 2 Classification of language Dryer (2013)

In the example E2, "*Q7*" in the first sentence and "*Q7*" in the second sentence is the exact same string.

6.1.2 Sub-string match

The Sub-string match feature is used to compare the surface forms of antecedent and anaphor. It means that one of them is a substring of another. For Example,

E3: "*The Home Ministry has asked Rajasthan, Gujarat, and Haryana to tighten security and deploy additional forces ahead of the verdict. The Ministry told the three States to beef up security and ensure that no violence takes place after the verdict."*
[The Hindu, April 25, 2018]

In the example E3 *"The Ministry"* is a substring of "*The Home Ministry"*

6.1.3 Word overlap

In the Word Overlap feature, both anaphor and antecedent share at least one word. For example

E4: "The Prime Minister of India leads the executive of the Government. The Prime Minister is the senior member of cabinet in the executive of government in a parliamentary system."
[https://en.wikipedia.org/wiki/Government_of_India]

In the example E4, the overlap between *"The Prime Minister of India"* and *"The Prime Minister"* exists.

6.1.4 Alias

Alias feature is used to check whether anaphor is an alias of antecedent or not. For example

> *E5: "Hundreds of Central American migrants from a caravan that crossed Mexico reunited in Tijuana on Wednesday and planned to cross the border together this weekend in defiance of threats by U.S. President Donald Trump to repel them. The timing of the migrant's arrival could compromise a flurry of talks this week to renegotiate the North American Free Trade Agreement (NAFTA), which Mr. Trump has repeatedly threatened to scrap if Mexico does not crack down on the flow of Central-Americans through its territory."*
> *[The Hindu, April 26, 2018]*

In the example E5, "*Mr. Trump*" is an alias of "*U.S. President Donald Trump"*.

6.1.5 Edit distance

Edit distance feature defines that the distance between both anaphor and antecedent is less than a specific number. It means that the number of predefined string operations (Insertion, Deletion, and Substitution) is used for transforming one head into the other. This feature is used in morphology rich languages like Bengali, German, etc. For example

> *German: Mann – Mannes*

To recognize inflectional suffixes, edit distance is less than or equal to 2.

6.2 Grammatical features

6.2.1 Definiteness

A noun phrase is considered as a definite if the head noun is changed by the possessive or demonstrative pronoun or even a definite article. For example

> *E6: "The chief minister's comments come in the backdrop of the protest by the residents of Keezhattur in Kannur against the national highway bypass through their paddy fields. He said the government was able to acquire and to a considerable extent between Kasaragod and Thalappady."*
> *[The Times of India, April 13, 2018]*

In this example, "*He*" in the second sentence refers to "*The chief minister"* of the first sentence. In *"The chief minister* "markable, the chief minister is modified with article "the". It is a definite Noun phrase.

6.2.2 Proper noun

The proper feature is used to indicate that one or both the anaphor and the antecedent are proper nouns. Proper nouns are also considered as antecedent. The names can be one of the most critical entities used in newspaper texts as well as in fiction. During the file-parsing, the names list filter out all the proper nouns that exist side by side in the text file. Probably that sequence shows the first and the last name of a person and shows the sports team, name of a company, institution, country, organization, etc. For example

E7: "*India is the world's largest democracy and according to UN estimates, its population is expected to overtake China's in 2028 to become the world's most populous nation.*"
[http://www.bbc.com/news/world-south-asia-12557384 accessed on 16 April 2018]

In this example E7, "*India*" is a proper noun. "*its*" refers to a proper noun "*India*".

6.2.3 Pronoun

The pronoun feature is used to indicate one or both anaphor and candidate antecedent as pronouns. The pronouns are also allowed to be possible antecedent candidates for other pronouns. Pronouns represent noun phrases because pronouns signify salience. And one pronoun can mention another pronoun that leads to a noun phrase. For example

E8: "*Ram took ABC medicine.* It$_1$ *is used to treat infection but harmful to take more doses of it$_2$.*"

In this example E8, "*it$_2$*" in the second sentence points to "It$_1$" also in the same sentence, and "It$_1$" points to "*ABC medicine*".

6.2.4 Grammatical relation

Both anaphora and potential antecedent must hold the same grammatical relation.

- Subject:

This is used to indicate whether one or both markables are subject or not. For example

E9: "*Kane Williamson has been a very good influence. He is very calm.*"
[Hindustan Times, April 27, 2018]

In this example E9, "*He"* is a subject which refers to *"Kane Williamson"* which is also a subject.

- Direct Object:

A direct object is used to indicate whether one or both markables are the objects or not. For Example

E10: "*Ram gave Sita a book. It was a nonfiction book.*"

In the example E10 "*It*" refers to "*a book*" which is a direct object.

- Indirect object:

An indirect object is used to check whether one or both the anaphora and the antecedent are indirect objects or not. Indirect objects are nouns or pronouns that state about ‘to whom’ or ‘for whom’ the action is performed by the verb and who is beneficiary of the direct object.

An oblique complement is a prepositional phrase or an adverbial phrase that behaves like a compliment. For example

E11: “*Ram gave Sita a book. It was a nonfiction book read and she likes it.*”

In this example E11, “*she*” in the second sentence points to “*Sita*” which is an indirect object in the first sentence.

- Syntactic Parallelism:

Syntactic parallelism is used to check that both the anaphor and the antecedent have the same syntactic role. For example

E12: “*The schoolboy took off the uniform after getting off the school bus. Heput off it andstarted playing.*”

In the example E12, “*He*”is a subject that refers to “*The schoolboy*” which is also the subject. Similarly, “*it*” refers to “*the uniform*” which is in the first sentence as well. This shows that anaphors commonly match their antecedents in their syntactic functions.

6.2.5 Number agreement

Number feature is used to find out that both the anaphora and the antecedent agree in number or not (both are singular or plural). For example

E13: “*Ayesha is playing Chess. She liked this game.*”

In the example E13, “*She*” in the second sentence points to “*Ayesha*” present in the sentence immediately before second sentence. Both “*She*” and “*Ayesha*” agrees in number. Both are singular.

6.2.6 Gender Agreement

Gender feature is used to find out that both the anaphora and the antecedent agree in gender or not. The grammatical gender of both anaphora and antecedent can have Masculine, Feminine, Neuter, or Either values.

The example E13 used in Number agreement also satisfies the gender agreement. Both anaphor “*She*” and antecedent “*Ayesha*” agree in gender. Both are representing feminine. “*She*” determines the gender of “*Ayesha*”.

6.2.7 Animacy

Animacy points to living entities. This feature determines whether both the anaphora and the antecedent are animated or not. Those anaphora and antecedent that belong to semantic class be it human or animal are animated, while others are not. For example

E14: "*The young boy got a car. He liked it.*"

In the example E14, to find the antecedent of "*He"* present in the second sentence, it is imperative to understand that "*The young boy*" is animated, and therefore, the antecedent of "*He"* is "*The young boy*".

6.2.8 Quotation

Quotation feature is used where texts have a significant number of direct speeches, such as newspaper texts. The position of both the anaphora and the antecedent is interdependent. If the anaphora has occupied a place within the quotation marks, then the antecedent would be in the quote too. Similarly, if the anaphora has not occupied a place within the quotation marks, then the antecedent would also be outside the quotations. For example

E15: "*Our players did their best by fulfilling the expectations of the country and won medal after medal," the PM said.*

[https://timesofindia.indiatimes.com/india/pm-modi-shares-his-mann-ki-baat-highlights/articleshow/63958344.cms]

In the example E15, "*their*" refers to "*Our players*", and both are presented in the quotation.

6.3 Semantic features

6.3.1 Semantic class

The semantic feature is used to check whether both the anaphor and the antecedent have the same semantic class or not. The semantic class can be identified by using a named entity recognizer (NER) or an external source such as WordNet. For example

E16: "*The young boy got a car. He liked it.*"

In the example E16, "*The young boy"* and "*He"* both belong to a person class.

6.3.2 Semantic relation

The anaphor and antecedent are related semantically; for example, one is a synonym or hypernym of the other one. This can be determined by an external source like WordNet.

- Lexical reiteration:

Lexical reiteration feature means that NP is reiterated twice or more within the same paragraph. It also includes reiterated synonym of noun phrases, which usually comes after the definite articles or demonstratives and a noun phrases sequence along with the identical head count termed as a lexical reiteration. For example

E17: "*Preparing the new bottle of toner. Tap five times on the bottle and rotate the bottle twenty times.*
Remove the lid from the bottle."
[http://files.oceusa.com/media/Assets/PDFs/TSS/external/VP2045_55_65/VP20x5DC_ConfigurationManual.pdf accessed on July 7, 2018]

In the example E17," *bottle of toner"* and "*the bottle*" are treated as a lexical reiteration.

- Collocation pattern:

The collocation patterns are like "noun phrase (pronoun), the verb" and "verb, noun phrase (pronoun)". This feature has true value when both anaphor and antecedent have identical collocation patterns. An example (Mitkov1998):

E18: *"Press the key down and turn the volume up… Press it again."*

In the example E18, it follows a verb-noun phrase and verb -pronoun pattern; therefore, "*it*" refers to "key".

- Indicating Verb:

In Indicating verb feature, NPs are next off a verb that belongs to a pre-determined set considered as an antecedent of anaphora. The predefined set is = {present, discuss, illustrate, summarize, identify, describe, examine, define, check, show, develop, report, review, outline, investigate, consider, explore, analyze, assess, synthesize, study, deal, survey, and cover}. If a verb belongs to the above-given set, then the first NP coming after it is considered as antecedent.

For example

E19: "*The Minister of National Economy of the Republic of Kazakhstan Timur Suleimenov reported on the plan for the implementation of enterprises and large companies of state and municipal property, work is being carried out on the privatization of Samruk-Kazyna facilities. In his turn, the Minister of Finance of the Republic of Kazakhstan Bakhyt Sultanov reported on the intermediate results of measures taken to accelerate the sale of state property and quasi-public sector and forthcoming tasks within the framework of the approved schedules. Since 2016, the Comprehensive Privatization Plan has implemented 221 facilities, 189 facilities are in the process of reorganization or liquidation.*"
[https://primeminister.kz/en/news accessed on April 30, 2018]

In the example E19, "*his*" refers to "*The Minister of National Economy of the Republic of Kazakhstan Timur Suleimenov*" because "*The Minister of National Economy of the*

Republic of Kazakhstan Timur Suleimenov" is first NP that immediately follows verb which reportedly belongs to a predefined set.

- Section Heading:

Section heading feature preferred NP present in the heading of the section that also appears in the ongoing sentence, which is a part of it. For example

E20: "*John Legend buries the hatchet with Kanye. They may "agree to disagree "over Donald Trump, but there is nothing but love between Kanye West and John Legend. ...."*
[Noida Times, April 30, 2018]

In the example E20, "*They*" refers to "*John Legend, Kanye*" that is present in the section heading.

- Non-prepositional noun phrases:

This feature describes a noun phrase that is not a part of the prepositional phrase that is preferred as an antecedent rather than the prepositional noun phrase. Another example (Mitkov1998):

E21: "*Insert the cassette into the VCR making sure it is suitable for the length of recording.*"

In this example E21, "*it*" refers to *"cassette*" because "the *cassette*" is the non-prepositional noun phrase.

- Nearest Noun Phrase (NNP):

NNP feature tells that the closest NP to the anaphor is used as an antecedent.

- Immediate reference:

"In constructions of the form
… you V1 NP… con (you) V2 it (con (you)) V3 it) where con ∈ {and/or/before/after…}", If a noun phrase comes next off V1 then it will probably be considered as a possible competitor for an antecedent of the pronoun "it" which comes next off V2. For example, taken from Holen (2007).

E22: "*To turn on the printer, press the Power button and hold it down for a moment."*

Here, in given an example, E22 "*it*" refers to "*the Power button*" as it satisfies pattern.

6.4 Positional features

6.4.1 Sentence distance

This feature indicates about the distance between the anaphor and the antecedent. The antecedent of some anaphors like reflexive pronoun is present in the same sentence. In complicated sentences, noun phrases in the previous clause are the best candidates for the antecedent of an anaphor in the succeeding clause, followed by noun phrases in the previous sentence, then by nouns situated two sentences additionally back and finally nouns three sentences further back. For anaphors in manageable sentences, noun phrases in the prior sentence are the most suitable candidate for antecedent, followed by noun phrases situated two sentences further back and finally nouns three sentences further back.

For example

E23: "*Jennifer does chores herself because she doesn't trust others to do them right.*"

[http://www.gingersoftware.com/content/grammar-rules/reflexive-pronouns/ accessed on April 04, 2018]

In this example E23, "*herself*" is a reflexive pronoun that refers to "*Jennifer*" and both are present in the same sentence.

6.4.2 Apposition

The apposition feature is used to check whether anaphor is in apposition to antecedent or not. For example

E24: "*PM Modi was received by Kong Xuanyou, Assistant Minister of the Ministry of foreign affairs of China, Luo Zhaohui (Ambassador), Tong Daochi (Vice-Governor of Hubei) and many others.*"
[The Tribune, Apr 26, 2018]

In the example E24," *Assistant Minister of the Ministry of foreign affairs of China Assistant Minister of the Ministry of foreign affairs of China*" is in apposition to "*Kong Xuanyou*".

6.4.3 Relative pronoun

The relative pronoun feature is used to find that the anaphor is a relative pronoun to the antecedent. For example

E25: "*Chadha, who is campaigning for labor rights, said it was ironic that on one hand, workers were dying in absence of basic needs while on the other, a lot of money collected for their welfare remained unspent.*"
[The Tribune, Apr 30, 2018]In this above example, "*who*" is a relative pronoun referring to "*Chadha*" as antecedent.

Table 1 summarizes the features used by different AR approaches in a different variation.

7 Feature Selection

Feature selection is a mechanism of selecting a small subset of the original feature set. Given a feature set ‘n’, this process is to find a feature subset of size ‘m’ where m<n. This provides feature sets that are non-redundant and relevant. To construct a feature vector for the anaphora resolution task, the ultimate goal is to find features that would help to solve the task. This section introduces feature selection methods that are classified into three groups.

7.1 Filter method

This method filters out features without considering the learning algorithm. To determine the relevancy of features, it uses a suitable ranking technique or selection evaluation function. The ranking method can be ‘Correlation Criteria’ or ‘Mutual Information’. While considering Mutual Information as a selection criterion, it selects the ‘k’ features with the highest mutual information. It assumes that features are highly correlated with the class. Mutual Information in Eq. (7) below incorporates three entropy measures (Chandrashekar and Sahin 2014):

- The entropy of X
- The entropy of Y and
- Their joined entropy.

$$\boldsymbol{I}(\boldsymbol{X};\boldsymbol{Y}) = \boldsymbol{H}(\boldsymbol{X}) + \boldsymbol{H}(\boldsymbol{Y}) - \boldsymbol{H}(\boldsymbol{X}|\mathbf{Y}) \tag{7}$$

The Mutual Information I (X; Y) defines that given Y is known, how much X is known; it is the reduction in unpredictability when Y is known.

7.2 Wrapper method

In this method, the feature subset is selected using the Induction Algorithm. This does not require knowledge of the algorithm. It selects a useful subset of features by considering the induction algorithm as an evaluation function. The primary goal is to use various unrelated feature sets and pick the one that provides the best estimation. This is a long-standing task.

7.3 Embedded method

This method is a mixture of the Filter Method and Wrapper Method and includes feature selection as a part of the training process. Mutual Information is also used in Embedded Methods. A typical example of embedded methods is numerous types of decision tree algorithms such as CART, C4.5, random forest (Sandri and Zuccolotto 2006), and other algorithms such as multinomial logistic regression and its variants (Cawley et al. 2007).

Table 1 Feature set

Features	Description
Lexical	
String match	As described above in Sects. 6.1.1, 6.1.2
Alias	As described above in Sect. 6.1.4
Edit distance	As described above in Sect. 6.1.5
Grammatical	
Definiteness	As described above in Sect. 6.2.1
Givenness/First Noun phrases	This feature has two parts: Subject and information (theme). In this category, it represents the subject of the sentence which is considered as antecedent
Proper Noun	As described above in Sect. 6.2.2
Boost Pronoun Indicator	As described above in Sect. 6.2.3
Head noun Emphasis	It describes any NP not contained in another NP
Existential Emphasis	It is predicate nominals in existential constructions
Accusative Emphasis	This feature tells about NPs that are direct objects
Non-adverbial emphasis	This feature describes that any NP not contained in an adverbial prepositional phrase
Indirect object and oblique component emphasis	As described above in Sect. 6.2.4
Number, Gender	As described above in Sects. 6.2.5, 6.2.6 respectively
Animacy	As described above in Sect. 6.2.7
Grammatical	
Syntactic parallelism	As described above in Sect. 6.2.4
Quotation	As described above in Sect. 6.2.8
Adjectival NP	This feature tells about that NPs which have adjectives modifying the head
Non-immediate clause	This feature tells about neither an immediate nor a current clause
N-Nom	It tells about nominative nouns
N-Poss	It tells about a possessive noun
N-Dat	It tells about dative nouns
Others	It tells about nouns with case suffixes
Relative pronoun	As described above in Sect. 6.4.3
Mention type	It defines mention types (name, nominal, or pronoun) of the anaphor and the antecedent. Additionally, it checks whether the anaphor is a definite pronoun or demonstrative pronoun or only a pronoun and checks whether every entity in the mention pairs indicates the proper name
Semantic	
Semantic class	As described above in Sect. 6.3.1.
Givenness (First Noun phrases)	The “given information” illustrated by NP in earlier sentences is the first NP in a non-imperative sentence if it is a good candidate for antecedents
Lexical reiteration	As described above in Sect. 6.3.2
Indicating verb	As described above in Sect. 6.3.2
Section heading preference	As described above in Sect. 6.3.2
Non-prepositional noun phrases	As described above in Sect. 6.3.2
Collocation pattern Preference	As described above in Sect. 6.3.2
Immediate reference	As described above in Sect. 6.3.2
Term preference	A noun phrase that shows a term in the field about that text

Table 1 (continued)

Features	Description
Frequent candidate	This feature means that three candidate noun phrases happen to take place most often in the sets of competing candidates for all pronouns
Punctuation preference	This feature gives the preference to the candidates which have a comma following them
Selectional restriction pattern	This feature describes the noun phrases existing in collocation with the verb preceding or following the anaphor
Sequential Instructions	A preference is given to NPs in the NP1 position of a structure of the form: "To V1 NP1… To V2 it…."
Positional	
Referential distance	As described above in Sect. 6.4.1
Markable distance	The distance between the two remarkable in terms of the number of mentions
Immediate clause	This feature tells about Clause next to the current clause
Non-immediate clause	This feature tells about neither an immediate nor a current clause
Appositive	As described above in Sect. 6.4.2

Some of the embedded methods carry out features weighting depends on regularization models. These methods are based on LASSO (Ma and Huang 2008) and Elastic Net (Zou and Hastie 2005).

For a more detailed overview of filter, wrapper, and embedded approaches, the following papers provide a useful reference (Chandrashekar and Sahin 2014; Aha and Bankert 1996; Kohavi and John 1997; Blum and Langley 1997).

8 Features selected in existing AR approaches

It is challenging for a machine to select a candidate antecedent to which anaphor refers due to the ambiguous nature of Natural language in the AR system. To make the AR system accomplish this task, it requires linguistic knowledge and features set. It also requires an annotated corpus for assessing features used in the resolution system. The use of suitable features has a great influence on the performance of the resolution system. Many features reported in the literature have been hand-crafted (features extracted from raw data using some specialized algorithm) and manually selected. An iterative procedure was used (Strube and Müller 2003) similar to the wrapper approach for feature selection as described in Wrappers for feature subset selection (Kohavi and John 1997) for determining the relevant feature set pronoun resolution in spoken dialogue. A single and multi-objective optimization for feature selection in anaphora resolution (Saha et al. 2011) was used based on the Genetic algorithm (Goldberg 1989). Sikdar et al. (2015a) developed feature selection techniques based on the differential evolution method for AR in a resource-poor language; particularly Bengali based on the single objective optimization and also did multi-objective optimization that is based on the principle of differential evolution (Sikdar et al. 2015b).

Evaluation metrics as discussed in the Sect. 5 for the AR system, are used to evaluate the performance of the AR system. Here, considering the importance of features for the

Table 2 Features used in various AR approach

Feature Used	Language	Mitkov (1998)	Holen (2007)	Sikdar et al. (2015a, b)	Hobbs (1978)	Lappin and Leass (1994)	Palomar et al. (2001)	Converse (2005)	Yang et al. (2006)	Mitkov et al. (2002)	Tanev and Mitkov (2002)	Chaves and Rino (2008)	Mutso (2008)	Küçük (2005)	Dakwale (2014)
Collocation pattern match or Semantic parallelism	Hindi, Estonian, French, German, Czech, Norwegian, Bulgarian	√	×	×	×	×	×	×	√	√	√	×	√	×	×
Semantic category/selectional/semantic restriction of noun	Hindi, Chinese, English, Bengali	×	×	√	×	×	×	√	√	×	×	×	×	×	√
First NP or Giveness	Estonian, French, German, Czech, Norwegian, Bulgarian, Portuguese, English, Arabic, Chinese, Turkish	√	√	×	×	√	×	×	√	√	√	√	√	√	×

Table 2 (continued)

Feature Used	Language	Mitkov (1998)	Holen (2007)	Sikdar et al. (2015a, b)	Hobbs (1978)	Lappin and Leass (1994)	Palomar et al. (2001)	Converse (2005)	Yang et al. (2006)	Mitkov et al. (2002)	Tanev and Mitkov (2002)	Chaves and Rino (2008)	Mutso (2008)	Küçük (2005)	Dakwale (2014)
Indefinite NPs/Definiteness	Estonian, French, German, Czech, Norwegian, Bulgarian, Portuguese, English, Italian, Japanese, Arabic, Chinese	√	√	×	×	×	√	×	√	√	√	√	×	×	×
Indicating Verb	Estonian, French, German, Czech, Norwegian, Bulgarian, Portuguese, Italian, Japanese, Spanish	√	×	×	×	×	√	×	×	√	√	×	√	×	×
Lexical Reiteration	Estonian, French, German, Czech, Norwegian, Bulgarian, Turkish, Bengali, English	√	×	×	×	×	√	×	×	√	√	√	√	√	×

Table 2 (continued)

Feature Used	Language	Mitkov (1998)	Holen (2007)	Sikdar et al. (2015a, b)	Hobbs (1978)	Lappin and Leass (1994)	Palomar et al. (2001)	Converse (2005)	Yang et al. (2006)	Mitkov et al. (2002)	Tanev and Mitkov (2002)	Chaves and Rino (2008)	Mutso (2008)	Küçük (2005)	Dakwale (2014)
Section Heading Preference	Estonian, French, German, Czech, Norwegian, Bulgarian	√	√	×	×	×	×	×	×	√	√	×	√	×	×
Prepositional Noun Phrases	English, Norwegian German, Portuguese, Bulgarian, Italian, Japanese, Spanish	√	√	×	×	×	√	×	×	√	√	√	×	×	×
Immediate Reference	Estonian, French, German, Czech, Norwegian	√	×	×	×	×	×	×	×	√	√	×	×	×	×
Sequential Instruction	Estonian, French, German, Czech, Norwegian	√	×	×	×	×	×	×	×	√	√	×	×	×	×

Table 2 (continued)

Feature Used	Language	Mitkov (1998)	Holen (2007)	Sikdar et al. (2015a, b)	Hobbs (1978)	Lappin and Leass (1994)	Palomar et al. (2001)	Converse (2005)	Yang et al. (2006)	Mitkov et al. (2002)	Tanev and Mitkov (2002)	Chaves and Rino (2008)	Mutso (2008)	Küçük (2005)	Dakwale (2014)
Referential Distance	Estonian, French, German, Czech, Norwegian, Turkish, English, Arabic, Chinese, Bengali, Portuguese, Italian, Japanese, Spanish	√	√	√	×	√	√	√	√	√	√	√	√	√	√
Term Preference	English, Portuguese, Bulgarian	√	×	×	×	×	×	×	×	√	√	×	×	×	×
Boost Pronoun	Estonian, French, German, Czech, Norwegian	√	√	×	×	×	×	×	×	√	×	×	√	×	×
Syntactic Parallelism	Estonian, French, German, Czech, Norwegian, Portuguese	×	√	×	×	×	×	√	√	√	×	√	√	×	×
Frequent Candidates	Estonian, French, German, Czech, Norwegian	√	×	×	×	×	×	×	×	√	×	×	×	×	×

Table 2 (continued)

Feature Used	Language	Mitkov (1998)	Holen (2007)	Sikdar et al. (2015a, b)	Hobbs (1978)	Lappin and Leass (1994)	Palomar et al. (2001)	Converse (2005)	Yang et al. (2006)	Mitkov et al. (2002)	Tanev and Mitkov (2002)	Chaves and Rino (2008)	Mutso (2008)	Küçük (2005)	Dakwale (2014)
Number, Gender, Person/Animacy/Type	Estonian, French, German, Czech, Norwegian, Chinese, English, Sanskrit, Bengali, Turkish, Italian, Japanese, Arabic, Spanish	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Frequent Candidates	Estonian, French, German, Czech, Norwegian	√	×	×	×	×	×	×	×	√	×	×	×	×	×
Declination indicator	Estonian, French, German, Czech, Norwegian	×	√	×	×	×	×	×	×	×	×	×			
Name indicator/Proper noun	Estonian, French, German, Czech, Norwegian, Bulgarian, English, Portuguese, Italian, Japanese, Spanish	×	√	×	×	×	√	×	×	×	√	×			

Table 2 (continued)

Feature Used	Language	Mitkov (1998)	Holen (2007)	Sikdar et al. (2015a, b)	Hobbs (1978)	Lappin and Leass (1994)	Palomar et al. (2001)	Converse (2005)	Yang et al. (2006)	Mitkov et al. (2002)	Tanev and Mitkov (2002)	Chaves and Rino (2008)	Mutso (2008)	Küçük (2005)	Dakwale (2014)
Quotation/Unquoted Text Preference	Estonian, French, German, Czech, Norwegian, Turkish, Bengali, English	×	×	√	×	×	×	×	×	×	×	×	√	√	×
Accusative emphasis (Direct object preference)	Estonian, French, German, Czech, Norwegian, English, Arabic, Chinese	×	√	×	×	√	×	×	√	√	×	×	×	×	×
Reflexive Pronoun Constraint	Turkish, Bengali, English, Arabic, Chinese, Hindi	×	×	√	×	×	×	×	√	×	×	×	×	√	√
Personal Pronoun Constraint	Turkish	×	×	×	×	×	×	×	×	×	×	×	×	√	×
Nominative Case Preference	Turkish, Bengali, English	×	×	×	×	×	×	×	×	×	×	×	×	√	×
Predicate Nominal Preference	Turkish, Bengali, English	×	×	×	×	×	×	×	×	×		×	×	√	×
Punctuation Preference	Turkish, Bengali, English	×	×	×	×	×	×	×	×	×	×	×	×	√	×

Table 2 (continued)

Feature Used	Language	Mitkov (1998)	Holen (2007)	Sikdar et al. (2015a, b)	Hobbs (1978)	Lappin and Leass (1994)	Palomar et al. (2001)	Converse (2005)	Yang et al. (2006)	Mitkov et al. (2002)	Tanev and Mitkov (2002)	Chaves and Rino (2008)	Mutso (2008)	Küçük (2005)	Dakwale (2014)
Antecedent of Zero Pronoun Preference	Turkish	×	×	×	×	×	×	×	×	×	×	×	×	√	×
Selectional restriction pattern	Bulgarian	×	×	×	×	×	×	×	×	×	√	×	×	×	×
Adjectival NPs	Bulgarian	×	×	×	×	×	×	×	×	×	√	×	×	×	×
Semantic features for the subjects of verbs	Chinese	×	×	×	×	×	×	√	√	×	×	×	×	×	×
The pronoun's headword and POS	Chinese	×	×	×	×	×	×	√	√	×	×	×	×	×	×
ClauseDept, TreeDepth	Chinese	×	×	×	×	×	×	√	×	×	×	×	×	×	×
Pronoun's sentence position	Chinese	×	×	×	×	×	×	√	×	×	×	×	×	×	√
Position pair	Chinese	×	×	×	×	×	×	√	×	×	×	×	×	×	×
Consecutive Sentence start	Chinese	×	×	×	×	×	×	√	×	×	×	×	×	×	×

Table 2 (continued)

Feature Used	Language	Mitkov (1998)	Holen (2007)	Sikdar et al. (2015a, b)	Hobbs (1978)	Lappin and Leass (1994)	Palomar et al. (2001)	Converse (2005)	Yang et al. (2006)	Mitkov et al. (2002)	Tanev and Mitkov (2002)	Chaves and Rino (2008)	Mutso (2008)	Küçük (2005)	Dakwale (2014)
VC pattern	Chinese	×	×	×	×	×	×	√	×	×	×	×	×	×	×
C commands	Chinese, English	√	×	×	√	√	×	√	×	√	×	×	×	√	×
ACE type	Chinese	×	×	×	×	×	×	√	×	×	×	×	×	×	×
Pragmatic features	Chinese	×	×	×	×	×	×	√	×	×	×	×	×	×	×
Closeness, NNP (Nearest NP)	English, Arabic, Chinese, Portuguese	×	×	×	×	×	×	×	√	×	√	√	×	×	×
Min-Expansion	English,Arabic, Chinese	×	×	×	×	×	×	×	√	×	×	×	×	×	×
Simple-Expansion	English, Arabic, Chinese	×	×	×	×	×	×	×	√	×	×	×	×	×	×
String Match	Bengali, English, Chinese	×	×	√	×	×	×	√	√	×	×	×	×	×	×
Markable distance	Bengali, English, Chinese	×	×	√	×	×	×	√	×	×	×	×	×	×	×
Appositive	Bengali, English, Spanish	×	×	√	×	×	√	×	×	×	×	×	×	×	×
String Kernel	Bengali, English	×	×	√	×	×	×	×	×	×	×	×	×	×	×
Mention Type	Bengali, English	×	×	√	×	×	×	×	×	×	×	×	×	×	×
LeftRight-Match	Bengali, English	×	×	√	×	×	×	×	×	×	×	×	×	×	×

AR task, it might happen that features that work well for a particular metric might not behave in the same way for some other metric. The performance of the system can vary significantly as most of it depends on the selected features and machine learning algorithm. One solution is to select features manually by looking into all available features combinations that can give the desired features set. But this solution is costly; hence it is advised to find an automatic method for the selection purpose.

Research in AR is categorized into three broad categories:

- Syntax-based approach
- Semantic-based approach
- Neural network-based approach

In this paper, the focus is on exploring various features used for anaphora resolution in different languages. Table 2 shows the features used in different languages by researchers to resolve anaphora in several approaches. Here,

√—denotes that the given approach utilizes the corresponding features to resolve anaphora.

×—denotes that the given approach did not utilize the respective feature to resolve anaphora.

8.1 Syntax-based approach

The logic behind the syntax-based approach is the set of rules, principles, and processes that govern the structure of the sentence. The word 'Syntax' alludes to the grammatical arrangement of words in a sentence and also refers to their relationship with each other. Syntax provides the syntactic or grammatical structure of a sentence, and identification of grammatical structure is made via syntactic or syntax tree that proves to be very helpful in finding the meaning of the sentence. Therefore, the syntax is vital information for the AR task and provides various additional information regarding the boundaries of sentences, clauses, and noun phrases (NP), etc. The syntax-based approach considers that the grammatical structures shown by a fully parsed syntactic tree exist. The parse tree provides antecedents by applying appropriate S*yntactic* and *Morphological constraints* on them.

Resolving pronoun references (Hobbs 1978) performed the first syntax-based approach for the AR task. He used syntactic features which are *Number, Person, Gender feature*, and *Selection constraint* selects the candidate antecedent of anaphora. The pronouns covered by this algorithm were '*he*', '*sh*e', '*it*', and '*they*' and were evaluated on 300 examples of pronoun occurrences in three different texts. The success rate reported was 88.3%. With a few *Selectional constraints*, the success rate rose to 91.7% (for cases where there were no multiple-choice or antecedent). In the cases where there were multiple antecedents to be chosen from, the success rate was reported to be 81.8%.

(Brennan et al. 1987) developed an algorithm based on centering theory (Grosz et al. 1995; Walker 1998) called BFP. BFP used syntactic features such as contra-indexing constraints (Reinhart 1976; Brennan et al. 1987) and morphological features like *Number* and *Gender* agreement to eliminate non- antecedent candidates. But it used centering principles to rank potential candidates. *Contra-indexing* constraints are explained with an example

E26: "*She does not like her.*"

In example E26, *"She" and "her"* cannot point to the same world entity; therefore *"She" and "her"* are contra-indexed. The disadvantage of BFP algorithm is that this algorithm is used to resolve inter-sentential anaphor only and difficult to handle intra-sentential anaphora resolution

Baldwin (1997) developed CogNIAC algorithm which is high precision pronoun resolution and exploited partial syntax tree for syntactic features. The author used *Number, Person, Gender features*, and *Selectional restriction*. The author evaluated on Narrative text, MUC-6. The Precision, Recall were 92%,64% on narrative text and 73%,75% on MUC-6 dataset respectively.

Walker (1989) developed an AR resolution algorithm that used discourse-based rule and syntactic constraints. The author evaluated three dataset samples consisting of *a chapter from Arthur Haley's novel Wheels, a Journalistic article from American weekly news magazine (July 7, 1975 edition of Newsweek)*, and one other. Those that are used by the Hobbs algorithm and others are of 5 human-human, keyboard-mediated, task-oriented dialogues. The overall success rate of this algorithm is 77.6%.

The RAP algorithm (Lappin and Leass 1994) implemented for German & English that is applied to the syntactic structure generated by McCord's Slot Grammar parser and used the *Morphological* filter such as *Number, Gender, Person, Syntactic filter*. In this feature, six conditions were used on Pronoun-NP Co-reference, salience measure that was inferred from the syntactic structure and are *Sentence recency, Subject emphasis, Head noun emphasis, Existential emphasis, Accusative emphasis, Non-adverbial emphasis, Indirect object, and oblique component emphasis*. A success rate of 86% was reported.

A new AR algorithm (Strube 1998) was developed based on discourse properties, and this algorithm called as S-list. This algorithm utilized features such as *Number, Gender agreements*, and *two types of syntactic features*: *Binding Constraints* (Chomsky 1981) and *Sortal Constraints*. *Binding constraints* are considered one of the numerous constraints that govern the resolution of pronouns. *Sortal constraints* are the ones which are ruled through semantics explained by the author (Elango 2005).

Tetreault (1999) developed Left-Right Centering Algorithm (LRC) to overcome the disadvantage of BFP that was explained by the author (Tetreault 1999). Another modification of LRC was done by (Tetreault 2001) and used some of the syntactic features such as *Number, Gender agreement, syntactic type, parent nodes, depth in the tree, position in utterance, presence or absence of a determiner, gender, coreference tag, utterance number, whether it was quoted, commanding verb, whether it was part of a title, whether it was reflexive, whether it was part of a possessive NP, whether it was in a prepended phrase, and whether it was part of a conjoined sentence* and *binding constraints.*

The Hobbs algorithm was further used by Ferrández and Peral (2000) for resolving zero-pronouns in Spanish texts, and the success rate was reported as 49.1%. Palomar et al. (2001) implemented the Hobbs algorithm and Lappin & Leass's RAP algorithm for resolving anaphora in Spanish texts and reported success rate was 62.7%, 67.4%, respectively.

Converse (2005) used *Gender, Number agreement, Selectional/Semantic restrictions (human, animate, abstract/concrete, location, and organization), Syntactic,* and *Pragmatic* feature. It has shown the impact of using a combination of these features on the AR system in the Chinese Language. The success rate of the algorithm on overt, third-person pronouns at the matrix level, for resolving matrix-level zero pronouns, on pronouns that appeared in subordinate constructions were reported as 77.6%, 73.3%, 43.3% respectively.

The Hobbs' Naïve Algorithm was further implemented by Tüfekçi and Kiliçaslan (2005) for Turkish by doing the modification and tested on ten toy sentences successfully. Han et al. (2006) implemented the naive version of the Hobbs algorithm and applied it to the two Penn Korean Treebank corpora. It was reported that the success rate of the algorithm was 46.1% and 62.2% for KTB1 (Penn Korean Treebank 1) and KTB 2 (Penn Korean Treebank2), respectively.

Yang and Tan (2006) proposed a method that is kernel-based and has parse trees information. To know the decision classifier and perform the resolution, it uses a S*tructured* feature (syntactic parse tree) and apply kernels to similar features with other typical features (syntactic information from parse trees). The features used were: *Reflexiveness, Type, Category, Subject, Object, Distance, First NP, Closeness,* and *Parallelism*. In this system, *Indefinite NP* is called *Category*, and *Distance* is called as *referential distance*. *NNP* is called as *Closeness*; *Reflexiveness* signifies that whether the pronominal anaphor is a type of reflexive pronoun. Here *Type* signifies that the pronominal anaphor falls into one of the following categories: Female-Person pronoun, or Male-Person pronoun, Single Gender-Neuter pronoun, or Plural Gender-Neuter pronoun. *Subject* feature specifies that either the candidate antecedent is in a subject position of a sentence, a subject position of a clause, or not in a subject position. *Object* feature specifies that either the candidate antecedent is in an object position of a verb, an object position of a preposition, or not in an object position. *Collocation pattern match* is called *Parallelism* in this case. Also, it includes *Structured Syntactic* feature which are three structured features containing different parse tree substructures: *Min-Expansion, Simple-Expansion,* and *Full-Expansion*. It used the ACE-2 V1.0 corpus that has two data sets, a Training and Development test (devtest). Training and devtest are further categorized into three categories: Newspaper (NPaper), Newswire (NWire), and Broadcast News (BNews). The success rate reported is indicated in Table 3 for normal features and for the combination as well. It was reported that when this system includes *structured* features with *normal* features then the success rate increases significantly by approximately 5%~8% for all the different domains.

The version of Hobbs' syntactic algorithm employed by de Arruda Santos and Carvalho (2007), resolve pronouns in the Portuguese language which is tested on three different corpora (Brazilian magazine, a literary book called O Alienista, Machado de Assis, Attorney General's Office of the Republic of Portugal).

Using Hobbs' algorithm and Lappin and Leass' algorithm it has been observed that in Brazilian magazine corpus, the success rate reported was 61.22% and 44.90% respectively. In literary corpus, the success rate reported was 49.68% and 33.91% respectively. In juridical corpus, the success rate reported was 40.40% and 35.15%, respectively.

A pronominal anaphora resolution algorithm (Murthy et al. 2007) was developed using salience measure and machine learning in the Tamil language. It considered only inter-sentential level. Features used by this algorithm are *Number, Person, Gender, Current sentence, Current clause, Immediate clause, Non-immediate clause, N-Nom, N-Nom, N-Poss, N-Dat, Others*. The *Current sentence* is a part of *Referential distance*. It was evaluated on a generic text taken from the CIIL (Central Institute of Indian Languages, Mysore, India) corpus. The overall Precision and Recall reported as 86.32% and 80.9%, respectively.

Linh (2007) developed a rule-based anaphora resolution approach and used on annotated Prague Dependency Treebank 2.0. This rule-based approach used the following features: *Subject, Subject in the main clause, Frequent noun, Collocation, Distance*. A Subject feature used as a *subject of a clause*, *subject in the main clause*, *frequent noun* indicates that nouns occur more than once in the current text. *Collocation* indicates that anaphor and antecedent have identical collocation patterns. The *distance* feature indicates that both the anaphor and the antecedent are belonging to the same sentence or different sentences. The F-measure of this system is 74.2%.

Table 3 Success rate of kernel-based method

Features	NWire (%)	NPaper (%)	BNews (%)
Normal Features	74.4	77.4	74.2
Normal features and Simple-Expansion	79.2	82.7	82.3
Normal features and Full-Expansion	81.5	83.2	81.5
Normal features and Min-Expansion	77.6	82.5	82.3

A modified variant of Hobbs' Naïve Algorithm for Hindi was proposed by Dutta et al. (2008). A heuristic-based anaphora resolution algorithm (Mahato et al. 2018) was developed for pronominal anaphora in Hindi Dialects. The authors considered first-person, second-person, and third-person pronouns for Hindi language and addressed the inter-sentential anaphora with considering 3–5 previous sentences before the current sentence and can be amended for more sentences as well. This algorithm resolves both Inter-sentential and Intra-sentential pronominal anaphora within the scope of three sentences. The features used by this algorithm are syntactic features such as *Number, Gender, Person, Headword of each sentence, Case marker* is a morphosyntactic feature which provides the role of a noun phrase*, Syntactic characteristics of an antecedent*. The authors used standard MUC evaluation metrics: precision, recall, and F-score to evaluate the algorithm. The Precision, Recall, and F-score reported were shown in Table 4.

Das et al. (2019) developed a rule-based anaphora resolution algorithm for the Bengali language. The author used the stem part and the POS of every word: person, noun, pronoun, compound noun, and object as features. The author has used 200 structured sentences from the Bengali text corpus, which is developed in TDIL (Technology Development for Indian Languages) project of the Government of India. The accuracy of this AR algorithm is 65.14%.

8.2 Semantic-based approach

A semantic-based approach is based on semantics in which the meaning is assigned to the symbols, characters, and words. The conceptual meaning of a word is used as Semantic features. This approach requires linguistic knowledge. Semantics are conditional.

The Lappin and Leass anaphora resolution algorithm (Lappin and Leass 1994) uses salience weight in deciding the antecedents to the anaphors. Kennedy and Boguraev (1996) proposed an extended version of the algorithm as described in (Lappin and Leass 1994). Full syntactic parsing is not required in this approach. This procedure is based on the outcome of a part of speech tagger, enhanced through the annotations of the grammatical role of each word of every sentence in the input text data. The success rate was reported as 75%.

Ge et al. (1998) developed a statistical approach for anaphora resolution that used features such as *Distance, Gender, Number, Animaticity, Governing head, Noun phrase repetition*. The success report was reported as 84.2% on Penn Wall Street Journal Tree-bank corpus.

Rich and LuperFoy (1988) used semantic knowledge with various features such as *Gender, Number, Recency* along with features about which entities are globally salient across entire utterance.

Palomar et al. (2001) proposed an algorithm for identifying third personal, demonstrative, reflexive, and omitted pronouns in unrestricted texts in the Spanish language. The

Table 4 Performance of heuristic-based anaphora resolution algorithm

Type of anaphora	Precision (%)	Recall (%)	F-score (%)
First person singular Pronoun	83	78	80
Second person Pronoun	84	70	77
Third person Pronoun	58	52	55

proposed algorithm used morphological agreement, syntactic conditions on NP-pronoun non-coreference, and preferences to resolve anaphora resolution in Spanish. It used feature *Intra-sentential, Inter-sentential*, which is part of the *Referential distance.Indefinite*, and*PNP, Indicating verb, Lexical reiteration*. *Lexical reiteration* is renamed as *AntRepeat, Proper noun feature* is used by the name *AntProper* and *Animacy* (*NoTime,*[23] *NoQuantity,*[24] *NoDirection,*[25] *NoAbstract,*[26] *NoCompany,*[27] *NPSentAnt*[28])which is used by RAP under name Head noun emphasis. The algorithm was evaluated in both literary texts and technical texts. The success rate reported as 76.8%. It was observed that without any semantic information, the success rate remains less than or equal to 75%.

Mitkov (1998) used a feature set for a resolution system which is called the Antecedent Indicators used by an AR system, as shown in Table 1. The features used in this system are *First noun phrase (FNP), Indefinite (INDEF), Indicating verb (IV), Lexical reiteration (REI), Section Heading (SH), Collocation pattern (CM), Non-prepositional noun phrases (PNP), Immediate reference (IR), Sequential instructions (SI), Referential distance (RD), Term preference(TP)*, and have assigned salience score to each feature. An extended version (Mitkov et al. 2002) was proposed in the original work by adding three new indicators viz, *Boost pronouns, Syntactic parallelism (SP),* and *Frequent Candidate (FC)*. The three new (proposed in 2002) and five original indicators proposed in Mitkov (1998) were implemented in different ways by MARS, the anaphora resolution system. Here, in the *Lexical reiteration*, it counts pronouns previously resolved to NP with the counting of an NP for all occurrences explicitly. The *collocation match* is implemented in such a manner that in the first step, for every verb that appeared in the document; both immediately preceding as well as immediately following heads of NP expression are written to file, and the immediately following NP expression is written in prepositions case. *FNP* was renamed to *obliqueness (OBL)*. *Referential distance (RD)* is used such that distance is calculated in terms of sentences, instead of clauses present in the sentences. It was reported that the MARS success rate was 59.35%. But, after applying the Genetic Algorithm, it increased to 61.55%.

Tanev and Mitkov (2002) developed a LINGUA which used the same feature set employed by Mitkov et al. (2002) and added three indicators according to Bulgarian language structure: *Selectional restriction pattern, Adjectival NPs,* and *Name preference*. *Selectional restriction pattern* is not the same as the *Collocation match* feature because it works on the complete range of '*Selectional restriction patterns*' which is related to a specific verb but does not match exactly with the lexical. The candidate is promoted to

[23] NP is not a time.

[24] NP is not a quantity.

[25] NP is not a direction.

[26] NP is not an abstract.

[27] NP is not a company.

[28] NP is included in another NP.

the antecedent if the verb preceding or following the anaphor is found to be in permissible collocation with a certain candidate. And it was reported that the success rate was 75% and improved by 2.20%, with the use of adjectival NP indicator. If this indicator applies to English, it would be intriguing to establish.

Liang and Wu (2004) developed anaphora resolution which employed features: *WordNet ontology, Number, Gender, Animacy, Syntactic parallelism, Semantic parallelism, Definiteness, Mention frequency, Sentence recency, Non-prepositional noun phrase,* and *Conjunction constraint*. *Conjunction constraint* is used to link words, phrases, and clauses. If the antecedent is connected with the anaphor by a conjunction, they can hardly have anaphora relation. This system is evaluated on Random text from Brown corpus. The success rate was reported as 77%.

Ho et al. (2004) developed the SmartINFO anaphora resolution system, which utilizes few *syntactic features* and *semantic features* rather than some syntactic dependency rules. Delmonte et al. (2006) developed an anaphora resolution system called GETARUNS, which used *grammatical relations* and *semantic roles* information for arguments and adjuncts. The author used a total of 75 *semantic features*, which were inferred from WordNet, and reward Human and Institution/Company labeled referring expressions. The F-measure was reported as 73.94% on Susanne Corpus.

Küçük (2005) used following features to resolve anaphora: *Number agreement, Reflexive pronoun constraint* requires that the nearest candidate to the pronoun must be the antecedent of a reflexive pronoun. *Personal pronoun constraint* necessitates that in a simple sentence, a personal pronoun cannot coexist with its antecedent. *Quoted/Unquoted text preference*, *Recency preference* is used by Mitkov (1998) under the name of Referential distance. *Nominative case preferences* are in which preference is given to the candidates who are in the nominative case. *Predicate nominal preference* is the one in which preference is given to a candidate if it is a predicate nominal. *Repetition Preference,* which is also used in Mitkov's based approach under name *Lexical reiteration*. *Punctuation preference* indicate that the preference is given to the candidates who have a comma following them. In *Antecedent of zero pronoun preference* feature, if a zero pronoun is considered, then the candidates having antecedents of zero pronouns in earlier sentences are given preference. The algorithm was evaluated on two different samples viz. the sample text from Metu Turkish Corpus and a Turkish child narrative. In the first experiment, Recall and Precision were 85.3% and 88%, respectively, and in the second experiment, Recall and Precision were 73.7% and 91%, respectively.

Holen (2007) has implemented factors in an Automatic Anaphora Resolution System for Norwegian (ARN) that is: *Number, Gender, Animacy*, and *those factors* were also used in Mitkov's Original Approach. RAP's *morphological filter* in (Lappin and Leass 1994), the Sentence *proximity factor* corresponds to *Sentence recency* salience weight in RAP, *Referential distance* factor in MARS, *Subject preference*, which has been called as *First noun phrases*. *Givenness* and *Obliqueness* used in both MOA and MARS respectively are the same as *Subject Preference* and the *First noun phrase*. *Boost Pronoun* is another feature used by the author in both MOA and MARS. This has also been implemented in Lappin &Leass RAP system under the name *Subject emphasis*. *Direct object preference* was represented in both RAP and MARS. In the RAP system, this was implemented under the name *Accusative emphasis (AE)*. The factor *Direct object preference* did not appear in MOA, but in MARS it did, along with the factors corresponding to ARN's factors viz.; *Indirect object preference, Subject preference,* and *Adverbial phrase penalization, a part of the Obliqueness factor*. *Indirect object preference* used in RAP is implemented just as a part of MARS as the *Obliqueness factor*. *Adverbial phrase penalization* was used by RAP as the name

Non-adverbial emphasis. *Prepositional phrase penalization*, which is used in MARS the impeding factor Prepositional NPs penalizes the NPs that are part of prepositional phrases. *Syntactic parallelism* is the factor that does not appear in either RAP or MOA, but the NPs with the same syntactical role as the anaphor has been given precedence by the MARS system. Section heading preference and *Indefiniteness penalization* factors are also applied by MOA and MARS. It was observed that the factors *Number, Gender, Animacy factor, Sentence proximity factor, Boost pronoun, Syntactic parallelism* are considered as hardcore of system. The subsequent two factors *Subject preference, Prepositional phrase penalization* were discarded, leaving five factors viz. *Direct object preference, Indirect object preference, Adverbial phrase penalization, Section heading preference, Indefiniteness penalization* in the system with some reservations. It was reported that by using all the seven factors, anaphora resolved 71.0% correctly. The final version of ARN contained all 5 factors by excluding the discarded ones, and this final version achieved 73.74% of correctly resolved anaphora when it was applied to the training corpus.

Chaves and Rino (2008) used Mitkov's indicators for Brazilian Portuguese. He changed Mitkov's basic algorithm to resolve the Third person pronoun resolution, which imparts NPs as antecedents. Mitkov categorized the factors into restrictive and preferential for identifying the antecedent candidates. To resolve anaphora, a *Restrictive facto*r gives a clue about the required properties of the antecedent candidates. It is recommended to discard the factors which are clueless about such properties; *Preferential factors* categorize all candidates as per their best chance of resolution rather than discarding. Mostly, *Preferential* and *Restrictive factors* are used. Mitkov's five antecedent indicators were included in RAPM, and three new interesting indicators were included. The indicators are as follows: *First NP, Lexical reiteration, Indefinite, Prepositional NP, and Referential distance (RD), Syntactic parallelism (SP), Nearest NP (NNP), Proper noun (PN)*. Some factors were used by RAPM by modifying a few indicators, such as repeated lexical items that are determined by using string matching. RAPM consider an NP as definite if its noun is changed by a definite article or demonstrative or possessive pronouns. For assessing RAPM with eight indicators, three corpora were used: A Law, A Literary, and A Newswire. The success rate was reported to be 67.01% for Newswire, 38.00% for Literary, and 54.00% for Law, respectively.

Mutso (2008) used the number factor to find out that the number of both anaphor and candidate matches, but in Estonian, there is no grammatical gender and the candidate noun match in number. Different indicators were used to identify candidate antecedent: *Giveness (FNP), Lexical reiteration (REI), Section heading (SH), Collocation pattern (CM), Referential distance (RD), Term preference (TP), Declination indicator, Name indicator, Indicative verbs, Syntactic parallelism, Quotation indicator, Boost pronoun*. Talking about all the indicators:

- *Giveness* feature was used in this system with slight changes like clauses were not annotated in corpora, a score would be assigned to the first noun phrase if the anaphora is also positioned in the same complex sentence. It did not prove to be much efficient, so this indicator was discarded from the system as Estonian is a free word order language.
- *Lexical Reiteration (REI)* is referred to as *Frequency Indicator*. In Estonian corpora, the word frequency was determined from one heading to another heading in text, unlike in the range of a paragraph that was initially used in Mitkov's algorithm.
- *Referential Distance (RD)*was implemented in a similar way as used by Mitkov in MARS. The only difference was in the factor score. If an anaphor lies in a complex

sentence and also there are no clauses located in the corpus, then a score is assigned to all the candidates of that sentence.

- *Section heading indicator (SH)*was also unable to give efficient results, so it was eliminated too.
- *The Indicative verb* was implemented in a similar way described originally by Mitkov but with some modification. It collected a list of verbs and phrases that commonly occur in newspaper texts before main entities.
- The author abolished *Syntactic parallelism* because of the non-promising results.
- The success rate of the resolution raised by about 0.5% leading to a success rate of 73.6% which is comparable to similar systems implemented for French, German, Czech, and Norwegian.

Sikdar et al. (2015a) used the following features to resolve anaphora resolution: *Sentence distance, String match, Markable distance Reflexive pronoun, First-person pronoun, Second person pronoun, Third-person pronoun, Number agreement, Alias, Semantic class, Appositive, Mention type, String kernel, LeftRight match. String match.*These were also implemented by Converse (2005). *Sentence distance* was also applied in Mitkov's approach under the name *Referential*. The *markable distance* was applied by Converse (2005) under the name *Minimum mention distance*. *First-person pronoun, Second person pronoun, Third-person pronoun* were also used by Küçük (2005) under *name Quoted/Unquoted text preference*, as *Quotation indicator* used by Mutso (2008). *Reflexive pronoun* was also used by Yang and Tan (2006), Küçük (2005). It was reported that for MUC, B3, CEAFM, CEAFE, and BLANC, F-measure values were 66.70%, 59.47%, 51.56%, 33.08%, and 72.75%, respectively.

Dakwale (2014) developed a hybrid approach to resolve Entity-pronoun references in Hindi and used following features viz. *Number* (Singular, Plural, Honorific), *Named Entity categories* (Person, Organization, Location), *Distance feature* (include number of chunks and number of sentences between the pronoun and candidate Noun phrase), *Animacy* (include human, animate, rest). It was reported that the rule-based system attained the accuracy of 60%, and the use of decision tree classifiers showed noticeable improvement, which is 10% over the rule-based system's accuracy. This concluded that semantic features like *Animacy* and *Named entity categories* give the important linguistic cue for anaphora resolution.

Abolohom and Omar (2015) proposed a method that employed both rule-based methods and learning-based methods for Arabic pronominal anaphora resolution. To check the significance and the impact of each feature, the authors performed a study on every feature and evaluated on the Arabic Quranic data set. It has been reported that this type of approach is more suitable and feasible for Arabic pronominal anaphora resolution. Features used by this approach are *Number, Gender, reflexive pronoun, anaphora nominative, Anaphora pronoun type* which are the third person singular pronoun, third person plural pronoun, *Subject, Frequency indicator (LR), Nearest NP, Non-Prepositional NP, FNP, Indefinite, Referential distance, Markable distance*. The Precision, Recall, and F-measure were reported as 71.7%, 78.2%, and 74.80%, respectively.

Kozlova et al. (2017) presented a hybrid method that employed both rule-based module and machine learning module for resolving anaphora in the Russian language. They employed the Extra Trees algorithm in the machine learning module and for handling the imbalanced data issue. Balance Cascade supervised algorithm was taken into action and used features were divided into five groups comprising of *Binary features, Non-binary categorical features, Numerical features, Numerical features derived from Word2vec,* and

Transposed vector. Binary features included morphological features, Non-binary categorical feature included grammatical features of anaphor and antecedent (number, gender, part of speech, case, type of anaphoric pronoun, animacy, syntactic relation), Numerical features derived from Word2vec include distance in words, sentences, nouns, verbs between anaphor and antecedent, and *Transposed vector was attained using SyntaxNet based on TF-IDF vectorization of morphological and syntactic tags* for both the anaphor and the antecedent in three directions of the syntactic context.

8.3 Neural network-based approach

As known that the AR system requires syntactic features, semantic features, and other deep knowledge to resolve anaphora, in the same context in the earlier days anaphora resolution system used to have handcrafted features, but nowadays the trend is changing from hand-crafted feature dependent methods to deep learning approaches which try to learn feature representation. With the origin of Deep Learning approaches, the goal is to minimize the need for hand-crafted features. Deep learning is simply a multilayered Artificial Neural Network (ANN).

Besides, the performance of deep learning is progressing due to the accessibility of additional data and more powerful computing machines such as Graphics Processing Units (GPU). Deep learning is a subset of Machine Learning, hence a family member of natural language processing. Words can be represented as vector space, which provides information about the semantic and syntactic structure. These vector models are called Word Embedding (Bengio et al. 2003; Mikolov et al. 2013a, b; Pennington et al. 2014) or also called as Type-level Word Embedding. This inspires to develop methods based on deep learning for AR. Recently trend is shifted from Type-level word embedding to Contextual word embedding (Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019), Embeddings from Language Models (ELMo) (Peters et al. 2018) which captures the contextual representation.

The Multi-Column Convolutional Neural Network-based approach described by Iida et al. (2016) is developed for Intra-sentential subject zero anaphora resolution. This uses features such as Pre-Trained 300-dimensional *word embedding vectors*, *Base* which includes a Japanese base phrase: bunsetsu (a phrase made up of minimum one content word after alternative function words), *Surface word sequence and Dependency tree of a targeted sentence* and *Predicate context*. It locally predicted the probability of zero anaphoric relation between all the feasible combinations of probable zero anaphor and antecedent regardless of the other relation between them in the corresponding sentence. It reported a high precision but a low F-score.

The Neural Network-based approach for anaphoric Zero pronoun resolution described by Chen and Ng (2016) was developed to address the two major issues in Zero anaphora resolution: the first being ‘the involvement of significant amount of the efforts for feature engineering for performance improvement’ and second ‘the inefficiency in utilizing *lexical features*’. This approach handled the above issues by considering ranking based on a deep neural network to rank the potential antecedents of an anaphor zero pronouns (AZP), which took feature vector representation of AZP, and n feature vectors representation for its n potential antecedents. All the vectors included *word embedding features* and *hand-crafted features*. In *word embedding features* for an AZP, it combined the embedding of its preceding word and its governing verb. It used word embedding of candidate antecedent’s head-word as embedding features for a candidate antecedent. The *hand-crafted features* included the positional, syntactic, and other relationships between an AZP and its potential antecedents. These features were like those used by (Zhao and Ng 2007; Kong and Zhou 2010; Chen and Ng 2013) as shown in Table 3. The performance of this system was assessed on the Chinese portion of the OntoNotes corpus to check the flexibility of the model for a language. This model attained excellent performance as the state-of-the-art results.

Liu et al. (2017) developed simple and novel approach to automatically generate large-scale pseudo training data for zero pronoun resolution. The authors utilized Attention based Neural Network (Hermann et al. 2015) and bidirectional Gated Recurrent Unit (GRU) (Cho et al. 2014) to reduce the dependency on handcrafted features. The F-score of this system on the Chinese portion of the OntoNotes 5.0 dataset was reported as 55.3%.

The deep neural network-based approach described by Yin et al. (2017a) to resolve Chinese Zero pronoun (ZP) anaphora aimed at representing Zero pronoun by its contextual information by capturing local and global information of candidate antecedent. The authors developed a method that has a ZP-centered long short- term memory network for representation of an anaphora zero pronouns (AZP) that are used to represent the preceding and succeeding contexts of the AZP separately for its representation. A two-level candidate encoder is used which consists of the local encoder (LE) that represents every candidate antecedent through a local vector representation by considering the content and contextual information of a single candidate. After this, these local vector representations are contemplated as input to global encoder which provides the global depiction for candidate antecedents. The authors employed a similar approach used by Clark and Manning (2016) to create local depictions for a candidate antecedent. The features used for this local encoder model included *word embedding of the headword, word embedding of the first word, word embedding of the last word, word embedding of two preceding words, word embedding of two following words and averaged word embedding of the five preceding context words, five succeeding context words, and all the contents words.*

Lee et al. (2017a) developed a Pointer Network model-based system for AR in the Korean language. In this paper, AR was considered a sequential mention connection problem. This method represents the *Syntactic* and *Semantic* features to solve the anaphora resolution problem with the bidirectional encoding of input tokens and decoded the chain form of anaphora by utilizing attention mechanism. Also, it did not require any handcrafting features to implement anaphora resolvers. They evaluated model on ETRI anaphora resolution dataset and also evaluated on English and Chinese dataset (SemEval-2010 dataset (Recasens et al. 2010), and CoNLL-2012 dataset (Pradhan et al. 2012) to prove the language dependency.

The Zero pronoun specific memory network proposed by Yin et al. (2017b) has the capability of creating vector representations for zero pronouns with additional information received from their contexts and candidate antecedents. This model took advantage of semantic information through these continuously distributed vectors while resolving zero pronouns. It was this model which generated the vector representation of the candidate antecedents called as external Memory and introduced attention mechanism to align dynamically more informative candidate from the external memory concerning given AZP and used them to create the representation of AZP. The feature set used by Chen and Ng (2016) were consolidated in the form of vector representation in the attention model for prediction of the correct antecedent of an AZP. The F-score of this system on the Chinese portion of the OntoNotes 5.0 dataset was reported as 53.6%.

Another algorithm (Yin et al. 2018a) which enhanced the earlier work shown in Chen and Ng (2016), Yin et al. (2017a, b), Liu et al. (2017) to combine local and global decision making by utilizing Deep Reinforcement learning agent and Policy-Based Deep

Table 5 Features used for Zero anaphora resolution (Zhao and Ng 2007; Chen and Ng 2013, 2016)

Feature category	Feature	Description
Syntactic feature for Zero pronoun	$P_{l\_}$ Is_ NP	P_l (left child nodes of P parse tree node is ancestor of immediate left word W_l of ZP) is a Noun phrase)
	P_r _Is_ VP	Pr (right child node of P parse tree node is ancestor of W_r) is a Verbal phrase
	$P_{l\_}$ Is_ NP and $P_{r\_}$ Is_ VP	P_l is an NP and P_r is a Verbal phrase
	P_ Is_ VP	P is a Verbal phrase
	IP-VP	There is a Verbal phrase node such as its parent node is an IP (sentence) node in the path from W_r to ceiling node C
	Has_ Ancestor _NP	V (highest VP node in the parse tree that is immediately to the right of Zero pronoun) has an ancestor NP
	Has _Ancestor _VP	V has an ancestor VP
	Has_ Ancestor_ CP	V has an ancestor CP which is clause headed by complementizers
Lexical Features for Zero Pronoun	Z Is_ZP	Z is a zero pronoun
	V Is_VP	V is the Verbal Phrase node following Z
	Wi	Wi (is a transitive verb, an intransitive verb or a preposition) is the ith word to right of z (if I is positive), or the ith to the left of z (if I is negative)
Other features for Zero pronoun	First gap	Indicates that Zero pronoun ZP is the first gap in the sentence
	Left comma	W_l is a comma
	Subject role	Zero pronoun is subject
	Clause	V is in a matrix clause, an independent clause, a subordinate clause, or none of the above
	Is_ In _Headline	Zero pronoun is in the headline of the text
	Ant_gov_verb	Antecedent is a subject whose governing verb is lexically similar to the z' s governing verb
Distance feature for antecedent	Dist _Sentence	Zero pronoun and Antecedent are in the same sentence, or not
	Dist_Segment	Zero pronoun and Antecedent are in the same segment or not
	Sibling _NP_ VP	Both Zero pronoun and Antecedents are siblings in parse tree
	Closest _NP	Antecedent is the closest preceding NP candidate to Zero pronoun

Table 5 (continued)

Feature category	Feature	Description
Syntactic feature for antecedent	A_ Has_ Anc _NP	Antecedent has an ancestor NP node
	A _Has_ Anc_ NP_ In_ IP	Antecedent has an ancestor NP node (descendant of A’s lowest ancestor IP node)
	A_ Has_ Anc_ VP	Antecedent has an ancestor VP node
	A _Has_ Anc_ VP_ In_ IP	Antecedent has an ancestor VP node (descendant of A’s lowest ancestor IP node)
	A _Has_ Anc_ CP	Antecedent has an ancestor CP node
	A_Grammatical_ Role	Antecedent’s grammatical role is subject, object, or others
	A_ Clause	Antecedent is in a matrix clause, an independent clause, a subordinate clause, or none of the above
	A _Is ADV	Antecedent is an adverbial NP
	A_ Is_ TMP	Antecedent is a temporal NP
	A_ Is_ Pronoun	Antecedent is a pronoun
	A_ Is_ NE	Antecedent is a named entity

Table 6 Performance of Neural Network-based on Zero anaphora resolution approach

References	Neural network architecture used	Language	F-score (%)
Iida et al. (2016)	Multi-column Convolutional Neural Network	Japanese	56.9
Chen and Ng (2016)	Feed forward Neural Network	Chinese	52
Liu et al. (2017)	Bidirectional Gated Recurrent Unit (GRU) +Attention mechanism	Chinese	54.9
Yin et al. (2017a)	Long Short term Memory Neural Network	Chinese	53.6
Yin et al. (2017b)	Long Short term Memory Neural Network and Recurrent Neural Network-based	Chinese	67.2
Li et al. (2017)	Pointer Network	Korean	83.05(MUC),83.51(CoNLL) on development data 78.86(MUC) 80.07 (CoNLL) on test data
Yin et al. (2018a)	Long Short term Memory Neural Network and Feedforward Neural Network	Chinese	55.3
Yin et al. (2018b)	Attention-based Neural Network	Chinese	57.3

Reinforcement learning algorithm. It is used to learn the policy of making coreference decisions for zero Pronoun-candidate antecedent pairs. The features set are: *Syntactical feature, Lexical feature, Distance feature, and other features* used by (Zhao and Ng 2007; Chen and Ng 2013, 2016) as summarized in Table 5 were consolidated into this model. The evaluation of this model was done on the Chinese portion of the OntoNotes 5.0 data set. The F-score of this system was reported as 67.2%.

Self-attention-based model for zero pronoun resolution described by Yin et al. (2018b) is a pair-wise model: modeling Zero Anaphora and modeling Candidate Antecedent. This model took advantage to effectively capture useful information from associative texts. The candidate antecedent encoder based on attention is used to model key parts of the noun phrases regarding the representative vector of zero anaphora. The features used by this system are *Syntactic features* described in Table 5, *Lexical features* such as the words surrounding zero pronouns and/or their part of speech (POS) tags. Considering z as a zero pronoun, then the words surrounding z and/or their POS tags, including w1, w-1, POS(w1), POS(w-1) + POS(w1), POS(w1)+POS(w2), POS(w−2)+POS(w−1), POS(w1)+POS(w2) +POS(w3), POS(w−1)+w1, and w−1+POS(w1). Whether w1 is a transitive verb, an intransitive verb, or a preposition; whether w−1 is a transitive verb without an object, *other features* shown in Table 5. The value of the F-score was reported as 57.3%. Table 6 shows the performance and neural network architecture used for zero anaphora resolution.

Annam et al. (2019) developed a new AR algorithm using neural networks (Binary Classification Multilayer Perceptron) in human-human dialogues for the Telugu language. The authors used an online shallow parser built by the LTRC center at IIIT Hyderabad with different features. The authors used six various features such as *Gender, Number* and *Person, Part-of-Plural, Speaker-Hearer,* and *generated 100-dimensional word embedding from word2vec*. The model was trained by authors using Telugu pages in Wikipedia and Andhrajyothi newspaper. They used new features such as Part-of-Plural, Speaker-Hearer, which were not discussed in the previous AR system. The author introduced the *Speaker-Hearer feature* because a considerable amount of first and second person was involved in conversations and assigned -1, 1 for these two actors for making the distinction between them. In the case of under-sampling, the Precision, Recall, and F1 were reported as 50.4%, 42.8%, 43.85%, respectively. For oversampling, Precision, Recall, and F1 were reported as 83.3%, 90.0%, and 86.0%, respectively. The author also analyzed the performance of the model by considering various features to find out the significance of every feature. Considering only *Gender* feature, the precision, recall, and F1 was reported as 80.09%, 86.8%, and 82.5%, respectively, which is highest. While considering other features such as *Number* and *Person, Part-of-Plural, Speaker-Hearer*, *generated 100-dimensional word embedding from the word2vec model.* This AR model has outperformed the recent state of the art in the anaphora resolution for Telugu language anaphora (Jonnalagadda and Mamidi 2015), which has 61.1%.

Allein et al. (2020) developed two models for anaphora resolution in the Dutch language. The first model is a *Binary Classification Model,* which solely predicts the correct die or dat. The binary classification model classifier utilized a Bidirectional Long-Short Term Memory (BiLSTM) neural network. 100-dimensional *word embedding* of sentences from the Word2Vec Skip-gram model (Mikolov et al. 2013a, b) is used as a feature and achieved 84.56% accuracy. The second model is the *Multitask Classification Model* used to predict the correct die or dat and *its part-of-speech tag* at the same time. The model used Bidirectional Long-Short Term Memory (BiLSTM) with 200-dimensional pre-trained *word embedding* for sentences with *immediate context* around the 'PREDICT' token as an extra input. The model achieved accuracy for die/dat prediction and part-of-speech prediction as 88.63%, 87.73%, respectively. This resulted that the multitask classification model increased the performance of die/dat prediction by 4% approximately with incorporating POS information and increasing word embedding dimension with increasing bidirectional LSTM layers.

Yang et al. (2020) presented anaphora resolution approach based on Multi-attention based capsule network for personal pronouns resolution in Uyghur language and Uyghur language is an agglutinative language. They used Independently Recurrent Neural Network (IndRNN) (Li et al. 2018) to learn semantic information, multi-attention mechanism, Capsule network (Sabour et al. 2017; Hinton et al. 2018). The authors utilized *pretrained word embeddings, Distance, Semantic features, POS tag of word, position of the anaphor and the candidate antecedents.* They evaluated the model on corpus which consist of raw text collected and annotated the text. There is no corpus available for Uyghur language due to the minority languages. The Precision, Recall, F-measure of this model were reported as 86.63%, 81.25%, 83.85%, respectively.

8.3.1 Gendered pronoun resolution

As discussed in Sect. 8.3, there is progress obtained from Type level word embedding to contextual word embedding. However, these developments, in the form of gender biasness, have also brought new challenges. Kurita et al. (2019) provide a comprehensive methodology into measuring gender bias in ELMo, while Zhao et al. (2019a) also address the error and presented two different methodologies to minimize the bias in the representations by Data Augmentation and Neutralization. As a result, the pronominal AR has become critical with gender bias for the CoNLL dataset, and AR has become popular with the sentence representation method using Word Embedding. Most of the methods used for addressing anaphora resolution do not adopt clustering strategies. The method used for pronominal AR in the case of gender-specific pronoun is either Binary Classification or Span-Ranking.

Zhang et al. (2019a) developed a pronoun resolution approach by incorporating contextual information and external human knowledge for a third personal pronoun (she, her, he, him, them, they, and it) and possessive pronoun (his, hers, its, their, theirs). The authors followed the same approach outlined by authors (Lee et al. 2017b) and utilized Bidirectional LSTM, an inner-span attention mechanism to obtain the contextual information. The knowledge attention module is used to process the external human knowledge. There are two types of knowledge used for this model i.e. *Linguistic Features* (Plurality, Animacy, and Gender) and *Selectional Preference* knowledge. The authors utilized simple FFNN and Softmax pruning to eliminate complexity before implementing a knowledge attention mechanism. The authors used 300-dimensional GloVe embeddings (Pennington and Manning 2014) and the ELMo (Peters et al. 2018) for word embedding and combined them to use as word embedding for words. This model is evaluated on the CoNLL-2012 shared task corpus. It was reported that the value of Precision, Recall, F-measure as 75.2%, 87.7%, 81.0%, respectively.

Zhang et al. (2019b) used the same methodology described by Zhang et al. (2019a) except the knowledge attention mechanism. The authors used knowledge representation for every triplet from knowledge graphs. Both approaches adopted by (Zhang et al. 2019a, Zhang et al. 2019b) adopted span representation methodology utilized by Lee et al. (2017b). Features used by authors are *Commonsense knowledge graph (OMCS)* (Singh 2002), *Medical concepts (Medical-KG)* provided in i2b2contest (Uzuner et al. 2010), *Plurality, Animacy*, and G*ende*r. They evaluated the model on CONLL 2012 and i2b2 dataset. The reported value of Precision, Recall, F-measure were 93.4%, 95.9%, 94.6%, respectively on i2b2 dataset and 73.6%, 76.4%, 74.9%, respectively on CONLL dataset.

Tenney et al. (2019) developed edge probing tasks that employ two-layer Multi-Layer Perceptron (MLP) with contextualized word representation such as ELMo and BERT (Devlin et al. 2019). This model used contextual information of the span representation using the attention-pooling mechanism described by Lee et al. (2017). On the other hand, Clark et al. (2019) did an in-depth analysis of BERT.

Emami et al. (2019) developed a KNOWREF corpus consisting of over 8000 annotated text passages with ambiguous pronominal anaphora and developed task-specific metric called consistency. It measures the extent to which a model uses the full context (as opposed to a surface cue) to make a coreference decision and utilizes BERT (Devlin et al. 2019) language model. The task-specific accuracy was reported as 0.61.

Google AI language team released Gender Ambiguous Pronoun (GAP) dataset (Webster et al. 2018) to address the gender bias and favoring masculine entities. (Webster et al. 2018) employed *syntactic structure* and *transformer models* to deal with gender bias pronoun resolution. Therefore, this dataset became popular because it is used by Kaggle's Gender Pronoun Resolution competition.

Attree (2019) proposed methodology for resolving gendered ambiguous pronouns on the Gendered Ambiguous Pronouns shared task, which was the winner of the Kaggle competition. The author used the Evidence Pooling (EP) module in which they used an attention pooling mechanism on the cluster predictions of coreference models along with a fine-tuned and a pronoun pooling methodology.

Ionita et al. (2019) developed a BERT-based methodology to deal with gender-balanced pronoun resolution. The authors used *embedding from BERT for named entities and pronoun, fine-tuned BERT classifier, Hand-crafted features such as pre-trained coreference resolution model predictions, Predictions of a Multi-Layered Perceptron trained with ELMo* (Peters et al. 2018) *embeddings, Syntactic roles of entities and Pronoun, Positional and frequency-based, Named entities predicted for entities, GAP heuristics* described in Webster et al. (2018). The F1 score was reported as 92%.

Liu (2019) developed a methodology for the Gendered Pronoun Resolution challenge, where the author used augmentation techniques and applied it to feature-based BERT. This system is evaluated on the GAP dataset.

Xu and Yang (2019) developed a system for Gendered Ambiguous Pronoun Resolution by using pre-trained BERT with Relational Graph Convolutional Network (R-GCN) to obtain syntactic information from embeddings and did not use any external prediction models.

Wang (2019) developed a system named MSnet (Mention Score Network), which is a neural network model using pre-trained BERT for Gendered Pronoun Resolution. The author used the attention mechanism to determine the contextual representation for the span of an entity and vector representation of semantic similarity between the pronoun and the entities. The author used a feature-based approach and the fine-tuning approach separately to train the MSnet model and evaluated on the GAP dataset.

Liu et al. (2019) presented an end-to-end differentiable model based on memory network (Weston et al. 2015) that involved the storage and accessing mentions for online text processing with efficient training by using both a supervised anaphora resolution along with supplementary language modeling called ad RefReader model. The main advantage of this model is to exploit the unlabeled text that was a different feature from previous work. The evaluation of the system was on GAP dataset and overall F1 for

Table 7 Anaphora resolution performance results

AR algorithm	Classification class	Scope of anaphora	Dataset	Languages	Metrics Used	Metrics Value			
						Success rate	Precision	Recall	F-measure
Hobbs (Hobbs, 1978)	Hand-crafted feature	Inter-sentential and intra-sentential anaphor (Pronouns: *he*, *she*, *it and they*)	WiLLiam Watson's Early Civilization in China, pp. 21–69, the first chapter of 2 Arthur Haley's novel Wheels, pp. 1–6, and the July 7, 1975 edition of Newsweek, pp. 13–19, beginning with the article 'A Ford in High Gear'.	English	Hobbs metric (success rate)	88.3% without selectional constraints 91.7% with selectional constraints	–	–	–
Walker (1989)	Hand-crafted feature	Inter-sentential and intra-sentential	Chapter of novel book and part of American weekly news magazine	English	Hobbs metric (success rate)	77.6%	–	–	–
RAP (Lappin and Leass, 1994)	Hand-crafted feature	Inter-sentential and intra-sentential anaphor	Computer manuals	English	Hobbs metric (success rate)	74% (inter-sentential) 89% (intra-sentential)	–	–	–

Table 7 (continued)

AR algorithm	Classification class	Scope of anaphora	Dataset	Languages	Metrics Used	Metrics Value			
						Success rate	Precision	Recall	F-measure
Kennedy and Boguraev (1996)	Hand-crafted feature	Inter-sentential and intra-sentential anaphor	Random text from Press releases, product announcements, news stories, magazine articles, and World Wide Web pages	English	Hobbs metric (success rate)	75%	–	–	–
CogNIAC (Baldwin, 1997)	Hand-crafted feature	Inter-sentential and intra-sentential anaphor	Narrative text, MUC-6	English	Precision, Recall	–	92% (Narrative text) 73% (MUC-6)	64% (Narrative text) 75% (MUC-6)	–
AR algorithm	Classification class	Scope of anaphora	Dataset	Languages	Metrics used	Metrics value			
						Success rate	Precision	Recall	F-measure
Ge et al. (1998)	Hand-crafted feature	Inter-sentential and intra-sentential anaphor	Small portion of the Penn Wall Street Journal Tree-bank	English	Hobbs metric (success rate)	82.9%	–	–	–

Table 7 (continued)

AR algorithm	Classification class	Scope of anaphora	Dataset	Languages	Metrics used	Metrics value			
						Success rate	Precision	Recall	F-measure
S-List (Strube, 1998)	Hand-crafted	Inter-sentential and intra-sentential	Three short stories by Hemingway and three articles from the New York Times, fictional texts.	English	Hobbs metric (success rate)	70% (overall)	–	–	–
(Mitkov, 1998)	Hand-crafted feature	Inter-sentential and intra-sentential anaphor	Random sample text from a technical manual	English, Polish, Arabic	Success rate (MUC) = Precision/ Recall	89.7% (English), 93.3% (Polish) and 95.2% (Arabic)	–	–	–
(Mitkov et al. 2002)	Hand-crafted feature	Inter-sentential and intra-sentential anaphor	Eight different files from the domains of computer hardware and software technical manuals		Hobbs metric (success rate)	59.35% (before applying Genetic algorithm) 61.55% (after applying Genetic algorithm)	–	–	–
Liang and Wu (2004)	Hand-crafted feature	Inter-sentential and intra-sentential anaphor	Random text from Brown corpus	English	Hobbs metric (success rate)	77%	–	–	–

Table 7 (continued)

AR algorithm	Classification class	Scope of anaphora	Dataset	Languages	Metrics used	Metrics value			
						Success rate	Precision	Recall	F-measure
SmartINFO (Ho et al. 2004)	Hand-crafted feature	Inter-sentential and intra -sentential anaphor (singular 3rd person pronouns)	Open domain resources such as children's novel	English	Hobbs metric (success rate)	76.7% (inter-sentential) 79.4% (intra-sentential)	–	–	–
Delmonte et al. (2006)	Hand-crafted feature	Inter-sentential and intra -sentential anaphor	Susanne corpus	English	Precision, Recall, F-score		62.7%	90.1%	73.94%
Sikdar et al. (2015a, b)	Hand-crafted feature	–	–	Bengali	MUC, B3, CEAFM, CEAFE, and BLANC	–	–	–	MUC: 66.70% B3: 59.47% CEAFM: 51.56%(CEAFM), 33.08%(CEAFE), 72.75% (BLANC)
Iida et al. (2016)	Learned feature	Intra-sentential (subject zero anaphora)	NAIST Text Corpus	Japanese	Precision, Recall, F-measure	–	70.4%	49.2%	56.9%

Table 7 (continued)

AR algorithm	Classification class	Scope of anaphora	Dataset	Languages	Metrics used	Metrics value			
						Success rate	Precision	Recall	F-measure
Lee et al. (2017)	Learned Feature	Inter-sentential and Intra-sentential	ETRI anaphora resolution dataset English portion of SemEval-2010 dataset, Chinese portion of CoNLL2012 dataset	Korean, English, Chinese	MUC, B-cube, and CEAF	–	–	–	For English MUC: 58.42% (SemEval-2010), 71.16%(CONLL-2012) CONLL: 62.33% (SemEval-2010), 71.20% (CONLL-2012) For Chinese (CONLL-2012) MUC: 55.21% CONLL: 61.56% For Korean (ETRI) MUC: 81.11% CONLL:82.02%
Yin et al. (2018a)	Learned feature	Intra-sentential (subject zero anaphora)	Chinese portion of the OntoNotes 5.0 data set.	Chinese	Recall, Precision, and F-score	–	–	–	57.3 (overall)

Masculine gender, F1 for Feminine gender, and the bias (the ratio of F1for Feminine to F1 for Masculine) were reported as 72.1%, 72.8%, 71.4%, 98%, respectively in case of RefReader model with pre-trained language model and coreference model prediction.

Chada (2019) presented an extractive question answering (QA) formulation of pronoun resolution task and also not required the knowledge of antecedents to provide an answer for the pronoun resolution task.

Kocijan et al. (2019) developed a language-model-based methodology for pronoun resolution on WikipediaCoREferences Masked (WIKICREM) dataset using the BERT language model. They developed the WIKICREM dataset containing 2.4 M samples of English Wikipedia text and fine-tuned the BERT language model using this dataset to show the improvement in methodology for pronoun resolution. They also evaluated this model on GAP test data.

9 Anaphora resolution performances

This section presents the results achieved by the methodologies described in Sect. 8. This section also highlights the improvements in the performance of the approaches and sets the state-of-the-art on AR to this date. Table 7 shows the performance result of AR approaches, and also indicates the scope of anaphora handled by all methodologies and metrics used for evaluating a system. As it is observed that AR approaches are not evaluated on the same dataset, therefore no benchmark exists for the evaluation system. First, Mitkov and Hallett (2007) compared five pronoun resolution algorithms: Kennedy and Boguraev (1996), CogNIAC (Baldwin 1997), Hobb algorithm, RAP algorithm on corpus contains technical manuals. They showed that the preprocessing tool should be perfect and experimental condition has an impact on the performance of AR. Hobb's algorithm was the first syntax-based algorithm that used the features having less computational cost.

Firstly, both Hobbs and LRC search for antecedents within the same sentence in which anaphor exists (intra-sententially) and then for previous sentences (other than the current sentence) in which anaphor exists (inter-sententially). Hobbs algorithm gives preference to intra-sententially first. Whereas, the BFP algorithm gives preference to inter-sentential as this algorithm did not use any method for considering intra-sentential processing. LRC algorithm was developed to overcome the disadvantage of the BFP algorithm. LRC has low computational overhead, and BFP has a high computational load.

These algorithms are adapted for various languages such as English, German, Chinese, Hindi, etc. Semantic-based Approach for anaphora resolution utilized semantic knowledge to improve the performance of the AR task. (Kennedy and Boguraev 1996; Rich and LuperFoy 1988; Ge et al. 1998; Ho et al. 2004; Liang and Wu 2004; Delmonte et al. 2006) are semantic-based AR algorithms utilizing semantic information developed initially for the English language. Delmonte et al. (2006) used grammatical relations and semantic roles information for arguments and adjuncts using WordNet. All these are using hand-crafted features.

Therefore, the Syntax-based AR approach, Semantic-based AR approach is classified under the hand-crafted feature category, and the Neural network-based approach comes under the learned feature category due to the use of deep learning techniques. Syntax-based AR approach requires less information and has a low computational cost; the Semantic-based AR approach has a higher computational cost than the syntax-based AR approach because of the requirement of more knowledge to gain semantic information.

Table 8 Performance of Gendered pronoun resolution algorithms on CONLL 2012 dataset

Algorithm	Neural network architecture	Scope of anaphora	Languages	Language model used	F1	Precision	Recall
Tenney et al. (2019)	Multi-Layer Perceptron (MLP)	Inter-sentential and intra-sentential	English	CoVeCoVe (McCann et al. 2017), ELMo, OpenAI GPT (Radford et al. 2018), and BERT	91.4%	–	–
Zhang et al. (2019a)	Bidirectional LSTM, inner-span attention mechanism+ FFNN	Inter-sentential and intra-sentential	English	GloVe embeddings+ ELMo	81.0%	75.2%	87.7%
Zhang et al. (2019b)	bidirectional LSTM (BiLSTM) + Attention mechanism	Inter-sentential and intra-sentential	English	GloVe embeddings+ ELMo	75.5%	75.9%	75.65%
Clark et al. (2019)	Attention-based probing mechanism	Inter-sentential and intra-sentential	English	BERT	65.0%	–	–

Table 9 Performance of Gendered pronoun resolution algorithms on GAP dataset

Algorithm	Architecture	Scope of anaphora	Language model used	F-score (Masculine)	F-score (Feminine)	Bias= Feminine/ Masculine	Overall	Logloss
Chada (2019) QA model	Logistic Regression and Span-wise Max Pooling Layer	Inter-sentential and intra-sentential	BERT	90.9%	89.5%	0.98%	90.2	0.32
Kocijan et al. (2019)	Transformer architecture	Inter-sentential and intra-sentential	BERT	76.4%	78.4%	1.03%	77.4%	
Wang (2019) MSnet	Attention mechanism + Feed-forward neural network	Inter-sentential and intra-sentential	BERT	–	–	–	–	0.32 (Training data) 0.2795 (Test data
Attree (2019)	multi-head attention and feedforward (FFN)	Inter-sentential and intra-sentential	BERT	94.0%	91.1%	0.97%	92.5%	0.317
Ionita et al. (2019)	MLP	Inter-sentential and intra-sentential	BERT	92.7%	90.0%	0.97%	91.4%	0.346
Liu (2019)	FFNN and Bidirectional LSTM	Inter-sentential and intra-sentential	BERT	91.6%	90.8%	0.99%	89.26%	.179
Liu et al. (2019)	Memory network architecture	Inter-sentential and intra-sentential	BERT	72.8%	71.4%	0.98%	72.1%	–
Xu and Yang (2019)	Relational graph convolutional Network (R-GCN)	Inter-sentential and intra-sentential	BERT	79.9%	81.1%	1.01%	80.3%	.493
Kocijan et al. (2019)	Transformer model	Inter-sentential and intra-sentential	BERT	76.7%	79.45%	1.04	78.0%	–
Webster et al. (2018)	Syntactic structure + Transformer models	Inter-sentential and intra-sentential	Transformer model	72.7%	69.25%	0.95%	70.95%	–

It is perceived that trend is shifting from handcrafted features to the learned representation of features to reduce the dependency of handcrafted features as it requires a lot of time and human resources. For this purpose, there is the advancement of using word embedding or contextual embedding, which captures syntactic and semantic information and helps to reduce the dependency of handcrafted features with the help of deep neural networks. Although deep neural networks can offer this exceptional accuracy, it comes at the expense of high computational complexity.

Iida et al. (2016) used Learned Features for intra-sentential zero anaphora for the Japanese language. Lee et al. (2017) used learned features for handling intra-sentential and inter sentential anaphora for the Korean language. Still, gender biasness exists in the dataset used for the AR task by different authors. To resolve this issue, the Google AI language team has recently released Gender Ambiguous Pronoun (GAP) dataset (Webster et al. 2018) to mitigate the gender biasness and favoring masculine entities because the gender biasness problem exists in CONLL 2012 dataset. Sun et al. (2019) outlined the recent in-depth study about identifying and minimizing gender biasness in NLP tasks such as Sentence Representation Methodologies, Anaphora Resolution, and those that can predict gender-biased predictions. Therefore, the GAP dataset is used by Kaggle's Gender Pronoun Resolution competition[29] to build a system to resolve gendered pronouns. Therefore, the pronoun resolution (AR) task has become popular due to Kaggle's Gender Pronoun Resolution competition and gender bias in both the CONLL dataset and the pre-trained word embedding model. There is no criterion available for this task, which can be considered to compare the performance of approaches because existing available AR approaches are not evaluated on the same dataset. Precision and Recall and F-measure metrics are used for evaluating gendered pronoun resolution.

Table 8 presents the results of the Gendered pronoun resolution algorithm in terms of Precision, Recall, F1 score that was evaluated on CONLL 2012. The results for the Gender pronoun resolution algorithm evaluated on the GAP dataset are displayed in Table 9.

It is observed that Tenney et al. (2019) got very high scores. In contrast, Zhang et al. (2019a) made the use of external knowledge which has shown the improvement in the baseline model and the approach used by Clark et al. (2019) has the worst performance because it is not designed for pronoun resolution task. It was also observed that biasness exists in data, and also few authors (Kurita et al. 2019; Zhao et al. 2019a; Clark et al. 2019) found the existence of bias in ELMo and BERT. The removal of bias has been accomplished by various methodologies, which are concluded by seeing the results produced by Kaggle competition in Gendered Pronoun Resolution. Recent work shows that it is better to use a sentence representation model for the AR task and make use of external knowledge features.

10 Discussions

Based on the observations, the study reveals that approaches for AR are divided into the following three categories: *Syntax, Semantic and Neural Network.* Table 10 summarizes the properties and disadvantages of the most well-known approaches (as discussed in Sect. 8).

[29] https://www.kaggle.com/c/gendered-pronoun-resolution.

Table 10 Properties and limitation of approaches

Approach	Properties	Limitation
Syntax-based approach	Depends on the syntactic parse tree Act as generally filters to remove inappropriate antecedents	Include the study of syntactic features only Require accurate parser Require a lot of human efforts and time
Semantic-based approach	Depends on the semantic features Improves the performance of AR by using semantic features	Includes the study of linguistic syntactic and semantic features Requires great human efforts and time
Neural network-based approach	Effective in representing the contextual information and semantics of larger sequences in text Continuous learning of feature causes better performance Need less number of hand-crafted features	Kneads large amount of data for training Needs a large amount of computational power

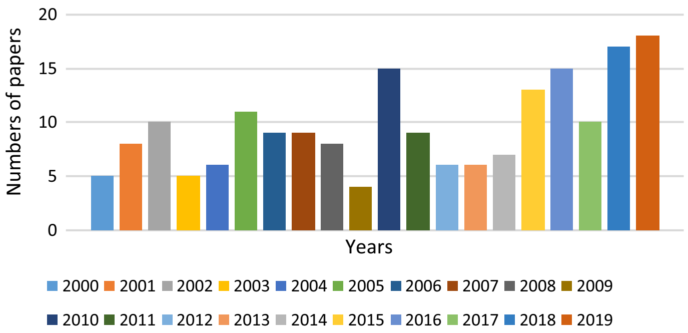

Fig. 3 Number of papers published per year

Syntax-Based approaches mostly considered the linguistic syntax-based features through a parse tree. Hobbs algorithm was the first syntax-based AR algorithm, but it assumed that the correct parser exists for sentence.

The *Semantic-Based* approach utilized semantic features for the AR process. Syntax and Semantic-based approaches are very time consuming due to the requirement of a massive amount of human effort to create the handwritten rule.

Neural Network approaches are used to capture the semantics of a long sequence of texts. Here, neural network-based approaches discussed in Sect.7.3 are using Deep Neural Network. Deep Neural Network is nothing but is an artificial neural network with multiple layers, and deep learning comes into picture that performs hierarchical learning of features. The deep neural network reduces the requirement of the number of handwritten rules/features. On the other hand, the neural network-based approach needs a large amount of data to exploit the computational power of CPUs/GPUs. Processing time is longer as compared to other traditional methods. But deep learning requires a large amount of annotated data, and the cost for creating such data is very high. For the AR task, there is a need for an annotated dataset; therefore, it is assumed that cost of annotation for a sentence is equal to the number of words in a sentence. Assuming that there are 'm' sentences, each has 'n' words; then the cost for annotation would be 'm*n'.

Figure 3 shows the number of papers published every year for anaphora resolution for different languages. It can be analyzed from the graph that in 2019 a higher number of papers were published from which it is perceived that the research work in the field of anaphora resolution is increasing. These papers used various features to solve the problems of anaphora in different languages. Figure 4 shows the frequency of features which are widely used by researchers to resolve the anaphora problem.

After analyzing Fig. 4, it is observed that Mitkov's approach to resolving anaphora has been used by different researchers for different languages. They have added different features as per their own language structure. Some researchers like (Broscheit et al. 2010) have used the BART system (Versley et al. 2008) to resolve anaphora that includes semantic and syntactic based features to improve the accuracy of anaphora resolution. A lot of effort is required in feature engineering for AR task; the goal is to reduce the effort for hand-crafted features. Nowadays, deep learning techniques are used to learn the feature representation. The deep neural network-based technique presented quality performance over traditional approaches.

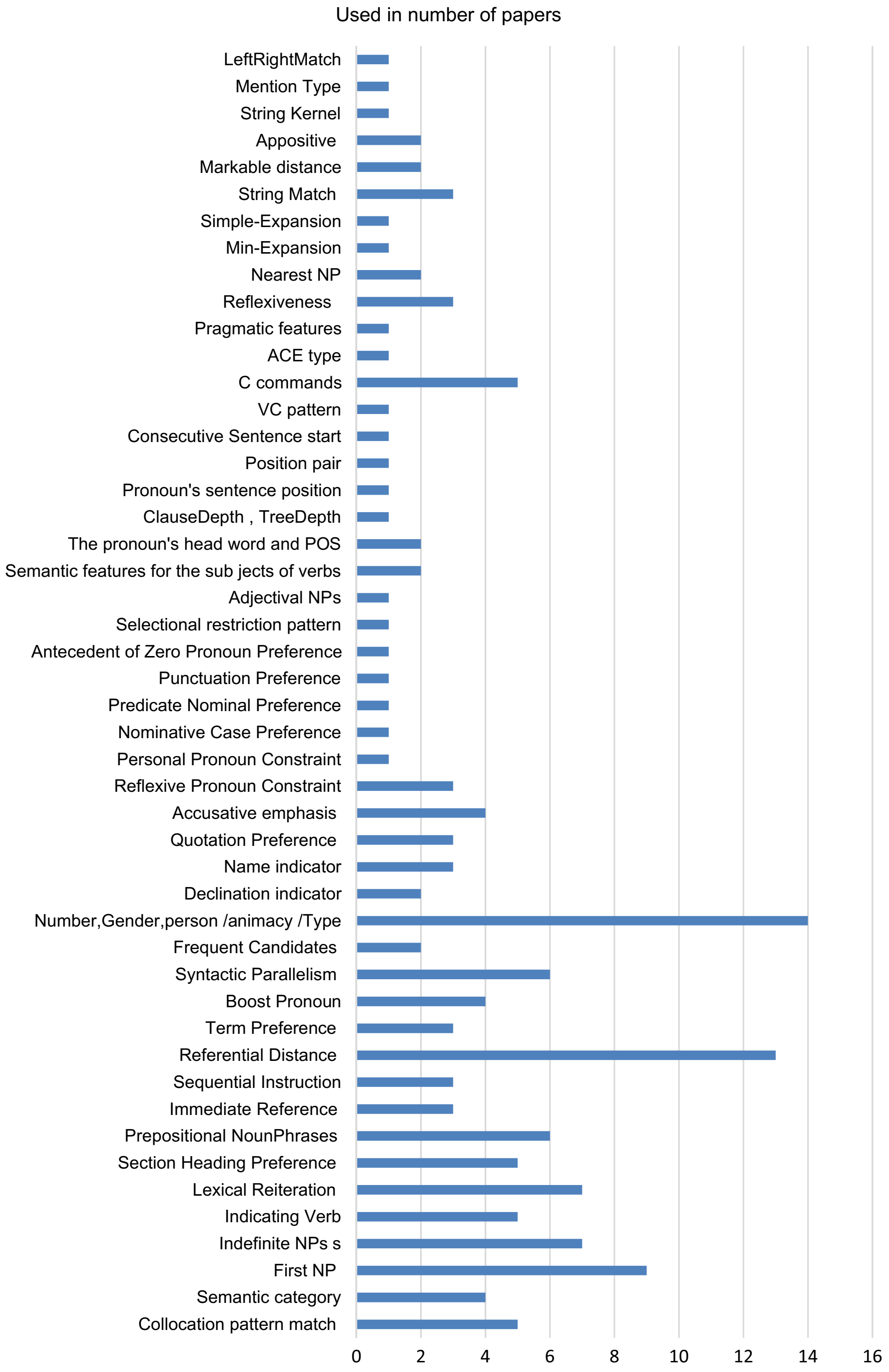

Fig. 4 Frequency of features used in AR approach

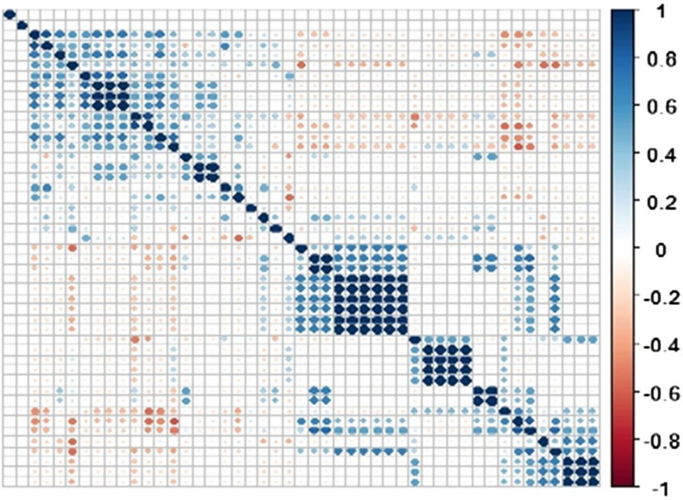

Fig. 5 Correlation matrix for features

The following answers are obtained against the identified RQs, as discussed in Introduction (Sect. 1), while having an in-depth exploration and analysis of the selected papers.

RQ1: What are the most relevant/sufficient/essential features for Anaphora Resolution approaches?

Solution: There are features which are necessary but not sufficient to remove incorrect antecedent. *Gender* is an essential feature that is used to eliminate the antecedents from the set of candidate antecedents. For example

E27: "*Ayesha has a small cat. It likes to eat.*"

In the above example E27, by applying this feature, due to the *Gender* disagreement with "*It*", "*Ayesha*" is removed from possible candidate antecedents {*Ayesha, a small cat*}, finally "*It*" points to "*a small cat*". But the main issue is when there are multiple antecedents which agree with gender feature, then what should be done.

To solve this issue, there is a need to impose some other Syntactic and Semantic Features. The *Numbe*r feature is an essential but not sufficient feature. It requires other features to resolve anaphor. *Person* feature is a necessary feature that is used by numerous AR methods like Lappin and Leass (1994). The antecedent and anaphora must agree in Person. The *Proximity feature* (distance between an anaphor and the antecedent) proved as a vital feature in earlier AR work. Still, this feature is not an ultimately decisive factor in AR in the Portuguese language (Nøklestad 2009). *Selectional restriction* feature is necessary because there are specific verbs that happen with living and non- living entities, respectively. This feature is used by many approaches like Ge et al. (1998). Some features depend on the type of corpus like newspaper text. *Quotation feature* is used for this, as explained in Sect. 3. As in some cases, *Morphological* features like *Number* and *Syntactic* features are not sufficient in the AR process; this enforces usage of essentially *Semantic* and *Pragmatic* features. The *semantic* features are essential, particularly for inter-sentential anaphora or indirect anaphora (Dutta et al. 2011). The *Grammatical* features are essential in word sense disambiguation; this feature can be used for the pronoun's resolution.

RQ2: How to extract and select efficient features adapting to different approaches used for anaphora resolution?

Solution: The relevant information from the data is deduced from feature extraction. The preprocessing such as tokenization, morphological analyzer, part of speech tagger, parser, named entity recognizer need to extract useful information and can use feature selection techniques to select efficient features used for different approaches. Wrapper, filter, embedding method is various feature selection methods to select relevant features. The feature selection method for selecting features in AR is explored in these papers (Saha et al. 2011; Sikdar et al. 2015a, b; Ekbal et al. 2011).

RQ3: Is there any correlation among features used in the AR system?

Solution: The correlation matrix derived for features using an R tool, as presented in Fig. 5 covers the answer to RQ3. Figure 5 shows the Pearson's correlation coefficient matrix of the attributes. Values close to -1 represent negative correlation, while a value near to 1 denotes the positive correlation. Here, Blue color indicates a positive correlation coefficient, and Orange indicates a negative correlation, and White color refers to the features that are not correlated to each other. With the help of this graph, a few highly correlated features can be located.

A perception from Fig. 5 is that the semantic-based highlights are progressively related to one another, and syntactic based highlights are increasingly connected with one another. Generally, it is unreliable to use correlation as a standard for feature selection as two correlated features can still improve AR tasks when they are in the same collection of features.

RQ4: Which features are languages specific that are used in the AR system?

Solution: From the research, it can be inferred that some features can be more vital than others, mostly due to the analyzed language characteristics. For example, in German Language, Edit distance and Grammatical gender are language-specific features. Positional features like Recency should be treated less significant than Syntactic features in the German newspaper, text shown by Wunsch (2006). Gender assignment is relatively complicated in the context of language structure in the German language shown by Nastase and Popescu (2009). For some groups of nouns, the derivation and assignment of Gender mostly depend upon the meaning and semantic rules. Basque language shown by Arregi et al. (2010a) has no Gender difference in its morphological system. Thus, it is not an accurate factor for finding an antecedent of a pronominal anaphora.

For the Spanish language, as shown by Chamorro (2018), it is required to consider *Grammatical gender, Elliptical subjects,* and *nouns* used attributively. There are many null subject languages like Italian, Arabic, Basque, Chinese, Hindi, and Korean. The behavior of features also depends on language characteristics. Like in the Hindi language, it is observed that the *Subject* and *Object* are important for anaphora resolution for reflexive and possessive pronouns. The definite article is not present in the Tamil language; thus, the *Definiteness* feature is not a determinative factor. The anaphora resolution task could be achievable and rely on the trustworthiness of language-specific features like *Person, Number,* and *Gender (PNG)* shown by Murthy et al. (2007). In a few languages like Arabic, Chinese, the *PNG agreement* between antecedent and anaphora is attained by analyzing the noun morphemes. For some languages such as Hindi language, verbal structures are analyzed for gaining *PNG* information. However, in some languages such as Japanese and Turkish, it is onerous to obtain *PNG* information because both verbal and noun phrase analyses are not sufficient; thus, there is a requirement of context-based or semantic analysis.

RQ5: What is the relationship between the feature and the approach used?

Solution: An AR approach must include an effective and appropriate set of features. Researchers consider that adequate universal features must be selected that is to be used in a different domain. The syntax-based approaches exploited only *Syntactic* features that utilized the syntactic structure of the sentence. Palomar et al. (2001) used a syntax-based approach to resolve anaphora in Spanish text and *Semantic* features were not included in this approach; thus, it is considerably less accurate. It can be concluded that an AR system requires *Semantic* information to obtain a success rate higher than 75%. Semantic-based approaches exploited both *Syntactic* and *Semantic* Features. Jain et al. (2004) utilized a vast range of *Syntactic, Semantic,* and *World knowledge-based* features for anaphora resolution. It was observed that the success rate with S*yntactic* features only was 85% and a combination of *Syntactic, Semantic* features was 98%, respectively. It is conjectured that *Semantic* features improve the accuracy of the resolution process.

As the traditional approaches (syntax-based, semantic-based approach) are based on hand-crafted features, Table 5 summarizes the set of features used by deep learning-based approaches for Zero pronoun resolution. Deep learning-based approaches used *Syntactic, Lexical,* and *other* features. The goal is to minimize the dependency on hand-crafted features and attempt to learn feature representations through the use of word embedding. It is observed that the performance of these approaches have improved over traditional approaches. It can also be perceived that bias exists in both data and contextual embedding model ELMo and BERT (Kurita et al. 2019; Zhao et al. 2019a; Clark et al. 2019). Kaggle offered challenge competition for Gendered Pronoun Resolution, because of this challenge; the progress has been attained in mitigating bias by using various techniques. On the other hand, the work on Pronoun resolution provides insightful results towards the better use of sentence representation models to the task and the use of external knowledge features. There is a lack of standard benchmark for the dataset and evaluation metrics for AR, as discussed in (Luo and Pradhan 2016). The GAP (Webster et al. 2018) dataset becomes a benchmark for gender bias evaluation.

As discussed in Sect. 2, there are many applications where the AR task has been used and the impact of AR tasks on applications such as Machine Translation, Information Extraction, Question answer system, etc. has been analyzed. Here, considering the most exciting application is a Machine Translation. There are few popular machine translators based on the neural network such as Google translator (Wu et al. 2016), Amazon translator[30] (Lu et al. 2018), DeepL,[31] Facebook translator[32] (Lample et al. 2018), Microsoft Translator[33] (Junczys-Dowmunt 2019).In these neural-based machine translators, it has not

[30] https://aws.amazon.com/translate/.

[31] https://www.deepl.com/translator.

[32] https://github.com/facebookresearch/UnsupervisedMT.

[33] https://translator.microsoft.com/help/languages/.

been explicitly specified as to how the authors have incorporated the AR task and not able to compare the performance in the context of the AR task.

10.1 Need of AR in text summarization

As discussed in Sect. 2, AR is one of the NLP's core component and has numerous possible downstream applications in NLP such as MT, IE, QA, Sentiment analysis, Text summarization, etc. The most exciting application of AR to us is a study of text summarization. We observed that there is less research work focusing on how AR can be successfully incorporated into a text summarization system and what issues it most likely can address. This section enumerates various scenarios and prominent approaches to text summarization.

Text summarization is used to summarize the text from news text, from internet information and summarize the reviews of movies, the product can buy from websites such as Amazon.com. Nowadays, customers make purchasing decisions for products based on customer's reviews, which are accessible on any online shopping site. Many customer's reviews can also be comprehensive and concise, and it is very time-consuming to analyze these reviews. Therefore, it is useful to summarize the lengthy reviews. While exploring the use of AR in text summarization, the case where AR could beneficial for the text summarization task is to determine the most appropriate term and check the coherence of the summary of the document produced by summarizer. The example of review (E28) has taken from Kaggle dataset[34] to make a clear picture of this task.

E28: (Amazon review)

Review1

I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than a processed meat and it smells better. My Labrador is finicky, and she appreciates this product better than most.

Summary: *Good Quality Dog Food*

The image in Fig. 6 shows the resolved references are returned after running ALLEN api[35](Gardner et al. 2018).

After running ALLEN api, resolved references are returned; In example E27, for example, all occurrences of elements of the anaphoric chain beginning with "*The product*" would be substituted by "*The product*", "*them all*" replaced with "*several of the Vitality canned dog food products*", and replaced "*she*" with "*My Labrador*". The resulting text would be as follows:

Review1

I have bought several of the Vitality canned dog food products and have found several of the Vitality canned dog food products to be of good quality. The product looks more like a stew than a processed meat and the product smells better. My Labrador is finicky and my Labrador appreciates the product better than most.

Summary: *Good Quality Dog Food*

[34] https://www.kaggle.com/currie32/summarizing-text-with-amazon-reviews#Summarizing-Text-with-Amazon-Reviews.

[35] https://demo.allennlp.org/coreference-resolution.

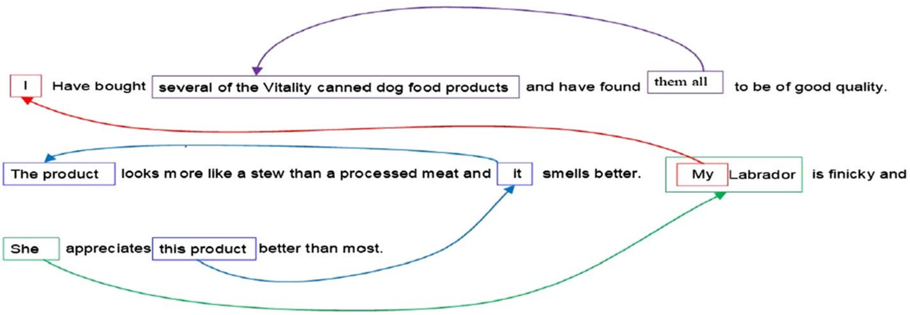

Fig. 6 Coreference resolution (ALLEN api)

AR will increase the frequencies of words referred by these pronouns and produces precise frequency counts. The main point is that repeatedly mentioned an entity is clearly essential. In E28, the main characters are "*My Labrador*" and "*food product*". Now the importance of AR has been established for text summarization. This section will provide an overview of the approaches which have been used for text summarization with AR.

The author (Sukthanker et al. 2020) explained the potential application of AR for fine-grained aspect-based sentiment analysis (Ma et al. 2018a, b). The use of AR helps derive multiple pronominal references to a particular aspect of the product, which in turn beneficial to extract the opinion associated with that specific aspect. Therefore, the use of AR information can also be useful in summarizing aspect-based sentiment summarization (Reis et al. 2014), Aspect-based Opinion Summarization (Mukherjee et al. 2020). Now the importance of AR has been established for text summarization. This section will provide an overview of the approaches which have been used for text summarization with AR.

Recent work for the text summarization (Xu et al. 2020) developed a discourse-aware neural summarization model based on BERT to capture the long-range dependencies in the text and also observed that some of the coherence issues are generated from missing or improper anaphora resolution

A recent approach for the text summarization (Antunes et al. 2018) in which a new method called Anaphoric Expression Solver (AES) was presented for extractive text summarization that endeavors to generate more coherent summaries by unraveling pronominal anaphoric expressions. The AES is based on Stanford CoreNLP[36] and evaluated on the CNN corpus (Lins et al. 2012). The AES method was able to achieve correct coreference substitution, improving the cohesion of equivalent to 81% of the total evaluated.

Durrett et al. (2016) aim to develop a single-document summarization system with compression and anaphoric constraints. They utilized Berkeley Entity Resolution System (Durrett and Klein 2014) and evaluated on New York Times Corpus (Sandhaus 2008) and RST Discourse Treebank (Carlson et al. 2003) with ROUGE (Lin 2004). They reported the value of ROUGE-1, ROUGE-2 was 42.2%, 25.9% respectively York Times Corpus and 26.3%, 8.0% respectively on RST Discourse Treebank.

Bayomi et al. (2016) aim to make the decision about whether AR can enhance the quality of produced summaries and assess whether AR has the same influence on text from

[36] https://stanfordnlp.git.hub.io/CoreNLP/.

Table 11 Performance of Text summarization with AR

Text summarization algorithm	AR algorithm used	Features implemented by AR	Languages	Performance without AR					Performance with AR					Corpus
				TF	TF*IDF	F-score (%)	Relative utility (%)	Cosine similarity (%)	TF	TF*IDF	F-score (%)	Relative utility (%)	Cosine similarity (%)	
Hassel (2000)	pronominal resolver for Swedish	Gender, Semantic information	Swedish, Danish, Norwegian, Spanish, French, English, German and Farsi	–	–	–	–	–	–	–	–	–	–	KTH News Corpus (Hassel 2001)

Table 11 (continued)

Text summarization algorithm	AR algorithm used	Features implemented by AR	Languages	Performance without AR					Performance with AR					Corpus
				TF	TF*IDF	F-score (%)	Relative utility (%)	Cosine similarity (%)	TF	TF*IDF	F-score (%)	Relative utility (%)	Cosine similarity (%)	
Kabadjov et al. (2005) (LSA-based text summarization)	GUITAR	Features utilized by MARS	English	–	–	For summarization ratio 15%:42 For summarization ratio 30%: 55.7	For summarization ratio 15%:59.5 For summarization ratio 30%: 64.5	For summarization ratio 15%:77.4 For summarization ratio 30%: 86.3	–	–	For summarization ratio 15%: 34.7(Substitution method for GUITAR 2.1) 44.1(Addition method for GUITAR 2.1) For summarization ratio 30%: 52.4 (Substitution method for GUITAR 2.1) 57.3 (Addition method for GUITAR 2.1)	For summarization ratio 15%: 53(Substitution method for GUITAR 2.1) 64(Addition method for GUITAR 2.1) For summarization ratio 30%: 62.6 (Substitution method for GUITAR 2.1) 67.8 (Addition method for GUITAR 2.1)	For summarization ratio 15%: 80.4(Substitution method for GUITAR 2.1) 80.5(Addition method for GUITAR 2.1) For summarization ratio 30%: 87.3 (Substitution method for GUITAR 2.1) 87.9 (Addition method for GUITAR 2.1)	CAST Corpus

Table 11 (continued)

Text summarization algorithm	AR algorithm used	Features implemented by AR	Languages	Performance without AR					Performance with AR					Corpus
				TF	TF*IDF	F-score (%)	Relative utility (%)	Cosine similarity (%)	TF	TF*IDF	F-score (%)	Relative utility (%)	Cosine similarity (%)	
Chen et al. (2005) (Mutlti-document summarization)	Not specified	–	Chinese	–	–	–	–	–	–	–	For summarization ratio 5%: 73.5 For summarization ratio 10%: 69.4	–	–	Chinese news text
Steinberger et al. (2007)	GUITAR 2.1 GUITAR 3.2	Features utilized by GUITAR algorithms as discussed in Section.	English	–	–	For summarization ratio 15%: 42.0 For summarization ratio 30%	For summarization ratio 15%: 59.5 For summarization ratio 30%	For summarization ratio 15%: 77.4 For summarization ratio 30%	–	–	On Cast corpus, For summarization ratio 15%: 34.7 (Substitution method) 44.1 (Addition method) For summarization ratio 30%: 52.4 (Substitution method) 57.3 (Addition method)	On Cast corpus, For summarization ratio 15%: 53.0 (Substitution method) 64.0 (Addition method) For summarization ratio 30%: 62.6 (Substitution method) 67.8 (Addition method)	On Cast corpus, For summarization ratio 15%: 80.4 (Substitution method) 80.5 (Addition method) For summarization ratio 30%: 87.3 (Substitution method) 87.9 (Addition method)	CAST, DUC-2002 corpus
Vodolazova et al. (2013)	RAP	Features used by RAP algorithm as discussed in Section	English	–	–	–	–	–	–	–	–	–	–	News articles (DUC-2002)

Table 11 (continued)

Text summarization algorithm	AR algorithm used	Features implemented by AR	Languages	Performance without AR					Performance with AR					Corpus
				TF	TF*IDF	F-score (%)	Relative utility (%)	Cosine similarity (%)	TF	TF*IDF	F-score (%)	Relative utility (%)	Cosine similarity (%)	
Bayomi et al. (2016)	Stanford Deterministic Coreference Resolution System (Lee et al. 2011)	Relaxed string match, Head word match, Semantic constraints, Distance, Alias, WordNet lexical chain, shallow discourse	English	–	–	–	–	–	–	–	–	–	–	Wikipedia article
Bayomi et al. (2016)	Stanford Deterministic Coreference Resolution System (Lee et al. 2011)	Relaxed string match, Head word match,Semantic constraints, Distance, Alias, WordNet lexical chain, shallow discourse	English	–	–	–	–	–	–	–	–	–	–	Wikipedia article
Durrett et al. (2016)	Berkeley Entity Resolution System	Mention type, string matching,head of a mention, The first word and last word of each mention, The word immediately preceding and the word immediately following a mention	English	–	–	–	–	–	–	–	–	–	–	New York Times Annotated Corpus

Table 11 (continued)

Text summarization algorithm	AR algorithm used	Features implemented by AR	Languages	Performance without AR					Performance with AR					Corpus
				TF	TF*IDF	F-score (%)	Relative utility (%)	Cosine similarity (%)	TF	TF*IDF	F-score (%)	Relative utility (%)	Cosine similarity (%)	
Durrett et al. (2016)	Berkeley Entity Resolution System	Mention length, in words, distance between mentions (sentence -level and number of mention (number of sentences and number of mentions)	English	–	–	–	–	–	–	–	–	–	–	New York Times Annotated Corpus
Antunes et al. (2018)	Stanford CoreNLP	Lexical, Syntactic, Semantic, and Discourse information	English	–	–	–	–	–		–	48.18(AutoS),[a] 47.18(AylienAPI)[b] 48. 95 (Classifier4J[c]) 46.48 [HP-UFPE FS Ferreira et al. (2014)]	–	66.68 (AutoS)[d], 64.43 (AylienAPI)[e] 64.97 (Classifier4J[f]) 64.04 (HP-UFPE FS (Ferreira et al. 2014))	CNN corpus (News articles)

[a] https://autosummarizer.com/

[b] https://docs.aylien.com/textapi/#language-support

[c] http://classifier4j.sourceforge.net/

[d] https://autosummarizer.com/

[e] https://docs.aylien.com/textapi/#language-support

[f] http://classifier4j.sourceforge.net/

different domains. This system utilized The Stanford Deterministic Coreference Resolution System (Lee et al. 2011) as the AR module. They used 70 Wikipedia abstracts from various subject domains viz. Politics, Accidents, Famous People, Natural Disasters, Sports, and Animals. They evaluated the system with AR and without the AR algorithm. They used characteristics such as *Readability and Understandability, Informativeness, conciseness, Overall quality of summary* for evaluating the summarization systemby using human evaluation. They reported a mean score, the variance, and p-value for every characteristic.

Vodolazova et al. (2013) also studied the impact of AR on extractive summarization system. The authors also investigated whether the performance of a text summarization approach is based on the topic of a document. The authors exploited anaphoric information produced by applying the RAP anaphora resolution algorithm in summarization task and evaluated on newswire articles collected from the Document Understanding Conference challenge of 2002 (DUC-2002). They used ROUGE-1 Recall on topics wise such as accidents, natural disasters, politics, sports, and famous people. The average ROUGE-1 Recall for the summarization system without AR, with AR were 38.9 and 39.3, respectively.

Steinberger et al. (2007) proposed a substitution approach for text summarization by using anaphoric information in Latent Semantic Analysis (LSA), and investigated that the performance boost of LSA-based summarizer is obtained on taking anaphoric information into account in text summarization. The authors utilized anahoric information to enhance the latent semantic representation of text and to determine whether the interpretation of anaphora in the extracted summary is similar to that in the original text. They utilized GUITAR anaphora resolution (Poesio and Kabadjov 2004; Kabadjov et al. 2005; Steinberger et al. 2005) and evaluated on CAST corpus (Orasan et al. 2003). and DUC-2002 corpus. Steinberger and Křišťan (2007) also used LSA based text summarization for multi documents with exploiting anaphoric information. They evaluated this system on Czech and English Corpora.

Orasan and St (2007) examined the effect of pronominal anaphora resolution on the performance of term-based summarization. The authors used three anaphora resolution algorithms: CogNIAC (Baldwin 1997), MARS (Mitkov 1998), and Kennedy and Boguraev (1996) for checking the effectiveness of AR on text summarization. The authors showed that a more accurate AR algorithm significantly increased the accuracy of a summarization approach. They also showed the success rate of these three AR approaches on the scientific domain (a corpus of journal articles collected from the Journal of Artificial Intelligence Research (JAIR)).

Steinberger et al. (2005) used GUITAR (Poesio and Kabadjov 2004) tool to improve the LSA-based text summarization system and evaluated on CAST project (Orasan et al. 2003). Kabadjov et al. (2005) done a case study that the use of anaphora resolution in the text summarization task could improve the performance of this task improve. They showed the performance of the summarization task with the use of perfect or automatic anaphora resolver. They utilized the GUITAR to exploit anaphoric information in LSA-based text summarization and also showed that the improvement in the performance of GUITAR is critical to achieving the substantial improvement in the performance of the summarization task. The author evaluated the system on the CAST corpus, DUC-2002 corpus[37] .

Another paper (Chen et al. 2005) has proposed the approach which utilized lexical chains for multi-document summarization of Chinese documents (Chinese newswire texts) and used the AR algorithm to enhance the fluency of the summary.

[37] https://www-nlpir.nist.gov/projects/duc/guidelines/2002.html.

Hassel (2000) exploited pronominal resolution module in text summarization for Swedish language. The author used natural natural or grammatical gender, and pluralized NPs for AR. They showed the impact of AR on text summarization.

After studying the text summarization approaches which exploit the anaphoric information, it has been found that existing AR algorithm such as GUITAR, RAP, Stanford Coreference resolution system, etc. are utilized in text summarization. The performance analysis of text summarization with AR and without AR is illustrated in Table 11. It has been observed from the previously mentioned research methodologies, the integration of the AR algorithm into text summarization is a daunting task. According to the fact, the current anaphora resolution algorithms are not flawless. Hence, if this algorithm is not incorporated correctly, then it might not boost the performance of the text summarization system. It will be complicated to improve the cohesion and local coherence without more robust anaphora resolution algorithm. In the future, research work for this area should emphasize on more in-depth evaluation on standard datasets for text summarization, multi-document summarization. Mostly, the focus of researchers is more on the extractive approach, it is imperative that the further work efforts are needed to propose and develop the effective summarization systems based on abstractive approach as well as hybrid approaches.

11 Challenges and future scope

Significant progress is seen in the field of the Anaphora Resolution over the past years. With the help of the Syntax-based approach, Semantic-based approach, and Neural Networks, several innovative ideas have been proposed by the researchers that have shown great success. Despite all the progress, still, some challenges are required to be solved. This section presents some of the challenges in the field of AR and some future research directions that will help in bringing new levels of study.

Language Dependent One can perform as much as possible linguistic analysis and design of NLP systems when not referring to a particular language such as Hindi, English, etc. This concept can be sought by modeling a natural language following the real-world notions, e.g., nouns and verbs abstracted as entities and actions, proper nouns as concrete entities (e.g., Mike) compared to abstract entities (e.g., boy), and so on. Hence, it is considered that the best method is to upgrade the current language modeling methodologies that are being utilized by all the majority of the tasks, along with anaphoric information.

Pre-Processing Time The accuracy of the pre-processing is yet too low. As a consequence, the performance of AR systems is still far from ideal.

Capsule Networks in AR Use of Capsule Networks is the latest technology in use in applications of NLP and is giving excellent results. Zhao et al. (2019b) presented a work "Towards Scalable and Reliable Capsule Networks for Challenging NLP Applications." In this, they evaluated the capsule network approach on two tasks: Multi-label Text Classification and Question Answering. Capsule networks helped them achieving remarkable results in both the fields in low resource settings with few training instances. The use of capsule networks in AR can bring standards.

Time Direction Independency A free text typically reflects the sequential series of events. Whereas, it does not always strictly hold, particularly in literature, where previous occurrences can be described later in the narration. However, a temporal order is generally followed in the discourse, and the AR technique has to consider that.

New Datasets Even though several large-scale datasets have been collected for common text classification tasks in recent years, still, there remains a need for new datasets for more innumerable challenging tasks such as QA with multi-step reasoning and text classification for multi-lingual documents. In the earlier datasets for AR task, there were issues of ambiguous pronouns (more than one candidate antecedent exists). The new datasets have the capability to overcome this issue too. There are some resource-poor languages (Indian languages such as Hindi, Bengali, etc.) for which a good set of benchmark dataset doesn't exist.

Computational Cost Despite a lot of research, the computational cost of an AR system is least considered. Most of the work done is performance specific. Many researchers did not consider computational costs during the approach used for this task; everyone has mainly focused on the performance of AR in terms of accuracy. Improvements can be made in AR systems in terms of computational cost, which can also lead us to some new results and observations.

12 Conclusions

The Anaphora resolution is an essential task for discourse. However, the use of Neural Network has shown great improvement. But due to the required word knowledge and inference problem that surrounds it, the AR task is considered as one of the most challenging tasks. The aim of this extensive literature search on Anaphora Resolution approaches is to explore and identify the features of AR, which has a high influence on the performance of the AR system and the trend shifting from the traditional approaches to deep learning approaches. In a deep learning approach, there is also a drift from Type-level word embedding to contextual word embedding.

The main findings of this review are outlined as follows:

- *Number, Gender, Person, Sentence recency/proximity, Selectional restriction, Animacy, Semantic features, World knowledge* is necessary features
- There are various techniques like Genetic algorithm, Multi-objective simulated annealing-based approach, and Differential evolution-based approach, which help feature selection in AR task.
- *Syntactic*- based features are related to each other, and *Semantic*-based features are related to each other.
- It is evident from the discussion in the literature that most of the features are shared across the languages for rule-based techniques. A few language-dependent features differ from system to system and language structure as well.
- The AR system requires Semantic features with Syntactic features for obtaining a success rate higher than 75%

For the AR task, a lot of effort is required for feature engineering. Nowadays, AR has been a changing trend from a hand-crafted feature dependent method to a deep learning-based method that automatically learns feature representation. The goal of this fieldwork has been to minimize the need for human/manually written features, which can be attributed to the origin of the fieldwork. There is also a requirement of the most appropriate method for evaluating performance for improving the AR task. The paper also did a discussion on AR tasks for different languages.

The Gap dataset (Webster et al. 2018) is considered as a benchmark dataset for the evaluation of the Gendered pronoun resolution task. The performance of present approaches significantly changes if one system is trained for one domain and adapted for another domain. The importance of improvements in the task can be found in applications such as Machine Translation (Bawden et al. 2018; Stojanovski and Fraser 2019; Voita et al. 2018, Phadke and Devane 2020), Named Entity Enrichment part of information extraction(Ting et al. 2019), Question Answering (Chowdhury and Chakraborty 2019), Recommendation system (Wohiduzzaman and Ismail 2018) etc.

Traditional methods have severe limitations in terms of computing resources and efficiency. Researchers have also cautioned that AR should be applied to NLP tasks only when it is necessary. Nevertheless, with technological advances, the availability of huge corpus, deep learning paradigms, new methodologies such as Attentive LSTM, Capsule network, has demonstrated promising results for the inclusion of AR modules in the real-world applications, thereby giving users a seamless experience.

Acknowledgements This work is supported by research grant from the Ministry of Education, Government of India. The authors would like to thank all reviewers for reviewing this work, and providing very insightful comments which helped us to improve the quality of manuscript. The authors also would like to thank the authors: Martin Hassel, Hercules Dalianis for providing detail about the Text summarization system developed by them after direct contact.

References

Abolohom A, Omar N (2015) A hybrid approach to pronominal anaphora resolution in Arabic. J Comput Sci 11(5):764

Aha DW, Bankert RL (1996) A comparative evaluation of sequential feature selection algorithms. Learning from data. Springer, New York, NY, pp 199–206

Allein L, Leeuwenberg A, Moens M-F (2020) Binary and multitask classification model for dutch anaphora resolution: die/dat prediction. ArXiv Preprint ArXiv:2001.02943. Retrieved from http://arxiv.org/abs/2001.02943

Annam V, Koditala N, Mamidi R (2019) Anaphora resolution in dialogue systems for south asian languages. ArXiv Preprint ArXiv:1911.09994. Retrieved from https://arxiv.org/abs/1911.09994

Antunes J, Lins RD, Lima R, Oliveira H, Riss M, Simske SJ (2018) Automatic cohesive summarization with pronominal anaphora resolution. Comput Speech Lang 52:141–164. https://doi.org/10.1016/j.csl.2018.05.004

Aone C, Bennett SW (1995) Evaluating automated and manual acquisition of anaphora resolution strategies. In: Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics, pp 122–129, https://doi.org/10.3115/981658.981675

Arregi O, Ceberio K, Díaz de Illarraza A, Goenaga I, Sierra B, Zelaia A (2010a) Determination of features for a machine learning approach to pronominal anaphora resolution in Basque, Procesamiento del Lenguaje Natural

Arregi O, Ceberio K, Díaz de Illarraza A, Goenaga I, Sierra B, Zelaia A (2010b) A first machine learning approach to pronominal anaphora resolution in basque. In: Ibero-American conference on artificial. Springer, Berlin, pp 234–243, https://doi.org/10.1007/978-3-642-16952-6_24

Attree S (2019) 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, Italy, Association for Computational Linguistics, pp 134–146

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

Baldwin B (1997) CogNIAC: high precision coreference with limited knowledge and linguistic resources. In: Proceedings of the ACL workshop on operational factors in practical, robust anaphora resolution for Unrestricted Text, Madrid, Spain, July 1997, pp 38–45, https://doi.org/10.1.1.45.1447

Batista-Navarro RT, Ananiadou S (2011) Building a coreference-annotated corpus from the domain of biochemistry. In: Proceedings of BioNLP 2011 workshop, association for computational linguistics, pp 83–91

Bawden R, Sennrich R, Birch A, Haddow B (2018) Evaluating discourse phenomena in neural machine translation. 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, Louisiana, Association for Computational Linguistics, pp 1304–1313

Bayomi M, Levacher K, Ghorab MR, Lavin P, O'Connor A, Lawless S(2016) Towards evaluating the impact of anaphora resolution on text summarisation from a human perspective. In: International conference on applications of natural language to information systems. Springer, Cham, pp 187–199

Bengio Y, Ducharme R, Vincent P, Jauvin C (2003) A neural probabilistic language model. J Mach Learn Res 3:1137–1155

Blum AL, Langley P (1997) Selection of relevant features and examples in machine learning. Artif Intell 97(1–2):245–271

Brennan SE, Friedman MW, Pollard CJ (1987) A centering approach to pronouns. In: Proceedings of the 25th annual meeting on association for computational linguistics, pp 155–162, https://doi.org/10.3115/981175.981197

Broscheit S, Poesio M, Ponzetto SP, Rodriguez KJ, Romano L, Uryupina O, Zanoli R (2010) BART: a multilingual anaphora resolution system. In: Proceedings of the 5th international workshop on semantic evaluation, Association for Computational Linguistics, pp 104–107

Carbonell JG, Brown RD (1988) Anaphora resolution: a multi-strategy approach. In: Coling budapest 1988 volume 1: international conference on computational linguistics

Carlson L, Marcu D, Okurowski ME (2003) Building a discourse-tagged corpus in the framework of rhetorical structure theory. Current and new directions in discourse and dialogue. Springer, Dordrecht, pp 85–112

Carter DM (1986) A shallow processing approach to anaphor resolution. Doctoral dissertation, University of Cambridge. https://doi.org/10.17863/CAM.740

Castagnola L (2002) Anaphora resolution for question answering. Doctoral dissertation, Massachusetts Institute of Technology

Castano J, Zhang J, Pustejovsky J (2002) Anaphora resolution in biomedical literature. In: Proceedings of the international symposium on reference resolution, Alicante, Spain

Cawley GC, Talbot NL, Girolami M (2007) Sparse multinomial logistic regression via bayesian l1 regularisation. In: Advances in neural information processing systems, pp 209–216

Chada R (2019) Gendered pronoun resolution using BERT and an extractive question answering formulation. In: Proceedings of the first workshop on gender bias in natural language processing, Florence, Italy, Association for Computational Linguistics, pp 126–133, https://doi.org/10.18653/v1/w19-3819

Chamorro G (2018) Offline interpretation of subject pronouns by native speakers of Spanish. Glossa: J Gener Linguist 3(1):27

Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28

Chatterji S, Dhar A, Barik B, Moumita PK, Sarkar S, Basu A (2011) Anaphora resolution for Bengali, Hindi, and Tamil using random tree algorithm in weka. In: Proceedings of the ICON-2011

Chaves AR, Rino LHM (2008) The mitkov algorithm for anaphora resolution in portuguese. In: International conference on computational processing of the Portuguese Language. Springer, Berlin, pp 51–60

Chen C, Ng V (2013) Chinese zero pronoun resolution: Some recent advances. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 1360–1365

Chen C, Ng V (2016) Chinese zero pronoun resolution with deep neural networks. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers), pp 778–788

Chen YM, Wang XL, Liu BQ (2005) Multi-document summarization based on lexical chains. In: 2005 international conference on machine learning and cybernetics (vol. 3), IEEE, pp 1937–1942

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1724–1734

Chomsky N (1981) Lectures on government and binding. Foris, Dordrecht. (1986a) Knowledge of language (Ch. 3), New York: Praeger. 1986b. Barriers. MIT Press, Cambridge

Chowdhury T, Chakraborty T (2019) CQASUMM: Building references for community question answering summarization corpora. In: Proceedings of the ACM india joint international conference on data science and management of data, pp 18–26

Clark HH (1975) Bridging. In: Theoretical issues in natural language processing, pp 169–174

Clark K, Manning CD (2016) Improving coreference resolution by learning entity-level distributed representations. In: Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 643–653

Clark K, Khandelwal U, Levy O, Manning CD (2019) What Does BERT Look at? An analysis of BERT's attention. In: Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP, pp 276–286

Cohen KB, Lanfranchi A, Corvey W, Baumgartner WA Jr, Roeder C, Ogren PV, Palmer M, Hunter L (2010) Annotation of all coreference in biomedical text: Guideline selection and adaptation. In: Proceedings of BioTxtM 2010: 2nd workshop on building and evaluating resources for biomedical text mining, pp 37–41

Converse SP (2005) Resolving pronominal references in Chinese with the Hobbs algorithm. In: Proceedings of the fourth SIGHAN workshop on Chinese language processing

Cooper JW, Kershenbaum A (2005) Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information. BMC Bioinform 6(1):143

Cuevas R, Paraboni I (2008) Using 'Low-cost'Learning Features for Pronoun Resolution. In: Proceedings of the 22nd Pacific Asia conference on language, information and computation, pp 377–383

Dakwale P (2014) Anaphora Resolution in Hindi. Doctoral dissertation, International Institute of Information Technology Hyderabad

Dariescu C, Gîfu D (2019) Ambiguous Interpretations in the legal discourse on both sides of the truth, Literature Mediator: Intersecting Discourses and Dialogues in a Multicultural World, https://old.upm.ro/ldmd/LDMD-06/Lds/Lds%2006%2039.pdf

Das A, Banerjee A, Maity S, Pal AR (2019) A rule based approach for anaphora resolution in bengali sentences. Int J Innov Technol Explor Eng 8(7):2652–2657

Davis E, Morgenstern L, Ortiz CL (2017) The first Winograd schema challenge at IJCAI-16. AI Magaz 38(3):97–98

de Arruda Santos DN, Carvalho AMBR (2007) Hobbs' algorithm for pronoun resolution in portuguese. In: Mexican international conference on artificial intelligence. Springer, Berlin, pp 966–974

Delmonte R, Bristot A, Boniforti MAP, Tonelli S (2006) Another evaluation of anaphora resolution algorithms and a comparison with GETARUNS' knowledge rich approach. In: Proceedings of the workshop on ROMAND 2006: robust methods in analysis of natural language data

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), Minneapolis, Minnesota, Association for computational linguistics, pp 4171–4186, https://doi.org/10.18653/v1/n19-1423

Do HH, Prasad PWC, Angelika M, Abeer A (2019) Deep learning for aspect-based sentiment analysis: a comparative review. Expert Syst Appl 118:272–299

Doddington GR, Mitchell A, Przybocki MA, Ramshaw LA, Strassel S, Weischedel RM (2004) The automatic content extraction (ACE) program tasks, data, and evaluation. In: LREC vol 21, pp 837–840

Dozier C, Zielund T (2004) Cross-document coreference resolution applications for people in the legal domain. In: Proceedings of the ACL-2004 workshop on reference resolution and its applications, barcelona, association for computational linguistics, pp 9–16

Dryer MS (2013) Determining dominant word order. The world atlas of language structures online. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available from http://wals.Info

Durrett G, Klein D (2014) A joint model for entity analysis: coreference, typing, and linking. Transactions of the Association for Computational Linguistics 2:477–490

Durrett G, Berg-Kirkpatrick T, Klein D (2016) Learning-based single-document summarization with compression and anaphoricity constraints. In:P roceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers), Berlin, Germany, Association for Computational Linguistics, pp 1998–2008. https://doi.org/10.18653/v1/p16-1188

Dutta K, Prakash N, Kaushik S (2008) Resolving pronominal anaphora in hindi using hobbs algorithm. Web J Form Comput Cognit Linguist 1(10):5607–5611

Dutta K, Prakash N, Kaushik S (2009) Application of pronominal divergence and anaphora resolution in English–Hindi machine translation. Polibits 39:55–58

Dutta K, Kaushik S, Prakash N (2011) Machine learning approach for the classification of demonstrative pronouns for Indirect Anaphora in Hindi News items. Prague Bull Math Linguist 95:33–50

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M (2005) The sequence ontology: a tool for the unification of genome annotations. Genome Biol 6(5):R44. https://doi.org/10.1186/gb-2005-6-5-r44

Ekbal A, Saha S, Uryupina O, Poesio M (2011) Multiobjective simulated annealing based approach for feature selection in anaphora resolution. Discourse anaphora and anaphor resolution colloquium. Springer, Berlin, pp 47–58

Elango P (2005) Coreference resolution: a survey. University of Wisconsin, Madison

Emami A, Trichelair P, Trischler A, Suleman K,Schulz H, Cheung JCK (2018) The hard-core coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. CoRR abs/1811.01747

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, pp 3952–3961, https://doi.org/10.18653/v1/p19-1386

Ferrández A, Peral J (2000) A computational approach to zero-pronouns in Spanish. In: Proceedings of the 38th annual meeting on association for computational linguistics, association for computational linguistics, pp 166–172

Ferreira R, Freitas F, de Souza Cabral L, Lins RD, Lima R, França G, Favaro L (2014) A context-based text summarization system. In: 2014 11th IAPR international workshop on document analysis systems, IEEE, pp 66–70

Gardner M, Grus J, Neumann M, Tafjord O, Dasigi P, Liu NF, Zettlemoyer L (2018) AllenNLP: a deep semantic natural language processing platform. In: Proceedings of workshop for NLP open source software (NLP-OSS), pp 1–6, https://doi.org/10.18653/v1/W18-2501

Gasperin C (2006) Semi-supervised anaphora resolution in biomedical texts. In: BioNLP'06: proceedings of the workshop on linking natural language processing and biology, association for computational linguistics, Morristown, pp 96–103

Gasperin C, Briscoe T (2008) Statistical anaphora resolution in biomedical texts. In: Proceedings of the 22nd international conference on computational linguistics-volume 1, association for computational linguistics, pp 257–264

Gasperin C, Karamanis N, Seal R (2007) Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In: Proceedings of DAARC, vol 2007

Ge N, Hale J, Charniak E (1998) A statistical approach to anaphora resolution. In: Sixth workshop on very large corpora, pp 161–170, https://doi.org/10.1.1.14.6342

Ghaddar A, Langlais P (2016) WikiCoref: an English coreference-annotated corpus of wikipedia articles. In: Proceedings of the 10th international conference on language resources and evaluation (LREC 2016), pp 136–142, Retrieved from http://www.lrec-conf.org/proceedings/lrec2016/pdf/192_Paper.pdf

Godfrey JJ, Holliman EC, McDaniel J (1992) SWITCHBOARD: telephone speech corpus for research and development. In: ICASSP-92: 1992 IEEE international conference on acoustics, speech, and signal processing, vol 1, IEEE Computer Society, pp 517–520

Goldberg DE (1989) Genetic algorithm. Search, optimization and machine learning. Addison-Wesley Longman Publishing Co, Inc. 75 Arlington Street, Suite 300 Boston, MA

Grosz BJ, Weinstein S, Joshi AK (1995) Centering: a framework for modelling the local coherence of discourse. Comput Linguist 21(2):203–225

Gupta A, Verma D, Pawar S, Patil S, Hingmire S, Palshikar G K, Bhattacharyya P(2018) Identifying participant mentions and resolving their coreferences in legal court judgements. In: International conference on text, speech, and dialogue. Springer, Cham, pp 153–162

Habimana O, Li Y, Li R, Gu X, Yu G (2020) Sentiment analysis using deep learning approaches: an overview. Sci China Inform Sci 63(1):1–36

Hahn U, Romacker M, Schulz S (2002) MEDSYNDIKATE—a natural language system for the extraction of medical information from findings reports. Int J Med Inform 67(1–3):63–74

Hajic J (1998) Building a syntactically annotated corpus: The prague dependency treebank. Issues of valency and meaning, pp 106–132

Halliday MK, Hasan R (1976) Cohesion in English. Longman, London, p 41

Hammami S, Belguith L, Ben Hamadou A (2009) Arabic anaphora resolution: corpora annotation with coreferential links. Int Arab J Inf Technol 6(5)

Han C, Han NR, Ko E-S, Palmer M, Yi H (2001). Penn Korean Treebank: Development and Evaluation. In: Proceedings of the 16th pacific Asia conference on language, information and computation, pp 69–78

Han NR, Prince EF, Palmer M (2006) Korean zero pronouns: analysis and resolution. University of Pennsylvania, Philadelphia

Hardmeier C, Federico M (2010) Modelling pronominal anaphora in statistical machine translation. In: IWSLT (International workshop on spoken language translation); Paris, France; December 2nd and 3rd, pp 283–289

Hassel M (2000) Pronominal resolution in automatic text summarisation (Master Thesis June 2000). DSV-Department of Computer and Systems Sciences, Stockholm University

Hassel M (2001) Internet as corpus—automatic construction of a Swedish News Corpus. KTH

He TY (2007) Coreference resolution on entities and events for hospital discharge summaries. Ph.D. thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science

Hermann KM, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P (2015) Teaching machines to read and comprehend. In: Advances in neural information processing systems, pp 1693–1701

Hinrichs E, Kübler S, Naumann K (2005) A unified representation for morphological, syntactic, semantic, and referential annotations. In: Proceedings of the workshop on frontiers in corpus annotations II: Pie in the Sky, pp 13–20

Hinton GE, Sabour S, Frosst N (2018) Matrix capsules with EM routing. In: International conference on learning representations

Ho H, Min K, Yeap W (2004) Pronominal anaphora resolution using a shallow meaning representation of sentences. PRICAI 2004 (Pacific Rim International Conference on Artificial Intelligence): trends in artificial intelligence. Springer, Berlin, pp 862–871

Hobbs JR (1978) Resolving pronoun references. Lingua 44(4):311–338. https://doi.org/10.1016/0024-3841(78)90006-2

Hobbs JR (1979) Coherence and coreference. Cognit Sci 3(1):67–90

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Holen GI (2007) Automatic anaphora resolution for Norwegian (ARN). In: Discourse anaphora and anaphor resolution colloquium. Springer, Berlin, pp 151–166

Hong KW, Park JC (2004) Anaphora resolution in text animation. In: Proceedings of the IASTED international conference on artificial intelligence and applications (AIA)

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, (June), pp 57–60

Hu Y, Guo Y, Liu J, Zhang H (2020) A hybrid method of coreference resolution in information security. Comput Mater Continua 64(2):1297–1315

Iida R, Komachi M, Inui K, Matsumoto Y (2007) Annotating a Japanese text corpus with predicate-argument and coreference relations. In: Proceedings of the linguistic annotation workshop, pp 132–139

Iida R, Torisawa K, Oh JH, Kruengkrai C, Kloetzer J (2016) Intra-sentential subject zero anaphora resolution using multi-column convolutional neural network. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 1244–1254

Ionita M, Kashnitsky Y, Krige K, Larin V, Logvinenko D, Atanasov A (2019) Resolving gendered ambiguous pronouns with BERT. In: Proceedings of the first workshop on gender bias in natural language processing, pp 113–119, https://doi.org/10.18653/v1/W19-3817

Jain P, Mital MR, Kumar S, Mukerjee A, Raina AM (2004) Anaphora resolution in multi-person dialogues. In: Proceedings of the 5th SIGdial workshop on discourse and dialogue at HLT-NAACL 2004

Jakob N, Gurevych I (2010) Using anaphora resolution to improve opinion target identification in movie reviews. In: Proceedings of the ACL 2010 conference short papers, association for computational linguistics, pp 263–268

Jonnalagadda HR, Mamidi R (2015) Resolution of pronominal anaphora for Telugu dialogues. In: Proceedings of the 12th international conference on natural language processing, pp 183–188

Ju TS, Roytberg A, Ladygina AA, Vasilyeva MD, Azerkovich IL, Kurzukov M, Grishina Y (2014) RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian. Komp’juternaja Lingvistika i Intellektual’nye Tehnologii, pp 681–694

Junczys-Dowmunt M (2019) Microsoft translator at WMT 2019: towards large-scale document-level neural machine translation. In: Proceedings of the fourth conference on machine translation (volume 2: shared task papers, Day 1(WMT 2019), pp 225–233

Kabadjov MA, Poesio M, Steinberger J (2005) Task-based evaluation of anaphora resolution: the case of summarization. In: Proceedings of the RANLP’05 workshop on crossing barriers in text summarization research, pp 18–25

Kameyama M (1997) Recognizing referential links: An information extraction perspective. In: Proceedings of a workshop on operational factors in practical, Robust Anaphora resolution for unrestricted texts, association for computational linguistics, pp 46–53

Karamanis N, Seal R, Lewin I, McQuilton P, Vlachos A, Gasperin C, Drysdale RA, Briscoe T (2008) Natural language processing in aid of FlyBase curators. BMC Bioinform 9(1):193

Kawahara D, Kurohashi S, Hasida K (2002) Construction of a Japanese Relevance-tagged Corpus. In: Proceedings of the 3rd international conference on language resources and evaluation (LREC-2002), Las Palmas, Canary Islands, pp 2008–2013

Kennedy C, Boguraev B (1996). Anaphora for everyone: pronominal anaphora resolution without a parser. In: Proceedings of the 16th conference on computational linguistics (COLING'96)-Volume 1, Association for computational linguistics, Copenhagen, Denmark, August 05–09, pp 113–118

Kilicoglu H, Rosemblat G, Fiszman M, Rindflesch TC (2016) Sortal anaphora resolution to enhance relation extraction from biomedical literature. BMC Bioinform 17(1):163

Kim J-D, Ohta T, Tateisi Y, Tsujii J (2003) GENIA corpus - A semantically annotated corpus for bio-textmining. Bioinformatics 19(SUPPL. 1):180–182. https://doi.org/10.1093/bioinformatics/btg1023

Kobayashi N, Iida R, Inui K, Matsumoto Y (2005) Opinion extraction using a learning-based anaphora resolution technique. In: Companion volume to the proceedings of conference including Posters/Demos and tutorial abstracts

Kobdani H (2012) A modular framework for coreference resolution. Dissertation, Stuttgart University, Germany

Kocijan V, Camburu O-M, Cretu A-M, 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), Association for computational linguistics pp 4294–4303, https://doi.org/10.18653/v1/D19-1439

Kocijan V, Lukasiewicz T, Davis E, Marcus G, Morgenstern L (2020) A review of winograd schema challenge datasets and approaches. arXiv preprint arXiv:2004.13831

Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1–2):273–324

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

Kong F, Zhou G (2010) A tree kernel-based unified framework for Chinese zero anaphora resolution. In: Proceedings of the 2010 conference on empirical methods in natural language processing, association for computational linguistics, pp 882–891

Kozlova A, Svischev A, Gureenkova O, Batura T (2017) A hybrid approach for anaphora resolution in the Russian language. In: 2017 Siberian symposium on data science and engineering (SSDSE), IEEE, pp 36–40

Kučera H, Francis WN (1967) Computational analysis of present-day American English. Dartmouth Publishing Group

Küçük D (2005) A knowledge-poor pronoun resolution system for Turkish, MSc Thesis, The Graduate School Of Natural And Applied Sciences, Middle East Technical University

Kuo JJ, Chen HH (2004) Event clustering on streaming news using co-reference chains and event words. In: Proceedings of the ACL-2004 workshop on reference resolution and its applications, association for computational linguistics, Barcelona, pp 17–23

Kurita K, Vyas N, Pareek A, Black AW, Tsvetkov Y (2019) Measuring bias in contextualized word representations. In: Proceedings of the first workshop on gender bias in natural language processing, pp 166–172, http://dx.doi.org/10.18653/v1/W19-3823

Lample G, Ott M, Conneau A, Denoyer L, Ranzato M (2018) Phrase-based & neural unsupervised machine translation. In: Proceedings of the 2018 conference on empirical methods in natural language processing, association for computational linguistics, pp 5039–5049, http://dx.doi.org/10.18653/v1/D18-1549

Lappin S, Leass HJ (1994) An algorithm for pronominal anaphora resolution. Comput Linguist 20(4):535–561, Retrieved from http://dl.acm.org/citation.cfm?id=203989

Lee H, Peirsman Y, Chang A, Chambers N, Surdeanu M, Jurafsky D (2011) Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In: Proceedings of 15th conference on computational natural language learning: shared task, association for computational linguistics, Stroudsburg, PA, USA, pp 28–34

Lee C, Jung S, Park CE (2017a) Anaphora resolution with pointer networks. Pattern Recogn Lett 95:1–7

Lee K, He L, Lewis M, Zettlemoyer L (2017b) End-to-end neural coreference resolution. In: Proceedings of the 2017 conference on empirical methods in natural language processing, association for computational linguistics, pp 188–197, http://dx.doi.org/10.18653/v1/D17-1018

Leech G, Garside R (1991) Running a grammar factory: The production of syntactically analysed corpora or treebanks. English computer corpora: selected papers and research guide, pp 15–32

Leidner JL (2008) Toponym resolution in text: annotation, evaluation and applications of spatial grounding of place names. Universal-Publishers, http://dissertation.com/book.php?book=1581123841&method=ISBN

Leidner JL, Sinclair G, Webber B (2003) Grounding spatial named entities for information extraction and question answering. In: Proceedings of the HLT-NAACL 2003 workshop on analysis of geographic references, association for computational linguistics, Morristown, pp 31–38, http://dx.doi.org/10.3115/1119394.1119399

Levesque H, Davis E, Morgenstern L (2012) The Winograd schema challenge. In: Proceedings of the thirteenth international conference on the principles of knowledge representation and reasoning, pp 552–561

Li S, Li W, Cook C, Zhu C, Gao Y (2018) Independently recurrent neural network (indrnn): building a longer and deeper rnn. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5457–5466

Liang T, Wu DS (2004) Automatic pronominal anaphora resolution in English texts. Int J Comput Linguist Chinese Language Process 9(1):21–40 **Special Issue on Selected Papers from ROCLING XV**

Lin CY (2004) Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out, pp 74–81

Linh NG (2007) Rule-based approach to pronominal anaphora resolution applied on the Prague Dependency Treebank 2.0 Data

Lins RD, Simske SJ, de Souza Cabral L, De Silva G, Lima R, Mello RF, Favaro L (2012). A multi-tool scheme for summarizing textual documents. In: Proceedings of 11st IADIS international conference WWW/INTERNET 2012, pp 1–8

Liu B (2019) Anonymized BERT: an augmentation approach to the gendered pronoun resolution challenge. In: Proceedings of the first workshop on gender bias in natural language processing, association for computational linguistics, pp 120–125, https://doi.org/10.18653/v1/W19-3818

Liu T, Cui Y, Yin Q, Zhang W, Wang S, Hu G (2017) Generating and exploiting large-scale pseudo training data for zero pronoun resolution. In: Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 102–111, https://doi.org/10.18653/v1/p17-1010

Liu F, Zettlemoyer L,Eisenstein J (2019) The referential reader: a recurrent entity network for anaphora resolution. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 5918–5925, https://doi.org/10.18653/v1/p19-1593

Loáiciga S, Wehrli E (2015) Rule-based pronominal anaphora treatment for machine translation. In: Proceedings of the second workshop on discourse in machine translation, pp 86–93

Lu Y, Keung P, Ladhak F, Bhardwaj V, Zhang S, Sun J (2018) A neural interlingua for multilingual machine translation. In: Proceedings of the third conference on machine translation: research papers, association for computational linguistics, pp 84–92, https://doi.org/10.18653/v1/w18-6309

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

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

Ma S, Huang J (2008) Penalized feature selection and classification in bioinformatics. Brief Bioinform 9(5):392–403

Ma Y, Peng H, Cambria E (2018a) Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In: AAAI, pp 5876–5883

Ma Y, Peng H, Khan T, Cambria E, Hussain A (2018b) Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis. Cognit Comput 10(4):639–650

Mahato S, Thomas A, Sahu N (2018) Heuristic algorithm for resolving pronominal anaphora in Hindi dialects. Advanced computing and intelligent engineering. Springer, Singapore, pp 41–51

Mahato S, Thomas A, Sahu N (2019) A relative study of factors and approaches for Hindi Anaphora resolution. Int J Manag IT Eng 7(12):176–188

Martin S, Poddar S, Upasani K (2020) MuDoCo: corpus for multidomain coreference resolution and referring expression generation. In: Proceedings of the 12th language resources and evaluation conference, pp 104–111

Matysiak I (2007) Information extraction systems and nominal anaphora analysis needs. In: Proceedings of the international multiconference on computer science and information technology, pp 183–192

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013a) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119

Mikolov T, Chen K, Corrado G, Dean J (2013b) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J (2020) Deep learning based text classification: a comprehensive review. arXiv preprint arXiv:2004.03705

Mitamura T, Nyberg E, Torrejon E, Svoboda D, Brunner A, Baker K (2002) Pronominal anaphora resolution in the kantoo multilingual machine translation system. In: Proceedings of the 9th international conference on theoretical and methodological issues in machine translation, Keihanna, Japan

Mitkov R (1995) Anaphora resolution in machine translation. In: Proceedings of the sixth international conference on theoretical and methodological issues in Machine Translation

Mitkov R (1998) Robust pronoun resolution with limited knowledge. In: Proceedings of the 18th international conference on computational linguistics, pp 869–875, https://doi.org/10.3115/980691.980712

Mitkov R (1999a) Introduction: special issue on anaphora resolution in machine translation and multilingual NLP. Mach Transl 14(3–4):159–161

Mitkov R (1999b) Anaphora resolution: the state of the art. School of Languages and European Studies, University of Wolverhampton, pp 1–34

Mitkov R (2001) Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. Appl Artif Intell 15(3):253–276

Mitkov R (2014) Anaphora resolution. Routledge

Mitkov R, Hallett C (2007) Comparing pronoun resolution algorithms. Comput Intell 23(2):262–297. https://doi.org/10.1111/j.1467-8640.2007.00305.x

Mitkov R, Schmidt P (1998) On the complexity of pronominal anaphora resolution in machine translation. Stud Funct Struct Linguist, pp 207–222

Mitkov R, Evans R, Orasan C (2002) A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In: International conference on intelligent text processing and computational linguistics. Springer, Berlin, pp 168–186

Mukherjee R, Peruri HC, Vishnu U, Goyal P, Bhattacharya S, Ganguly N (2020) Read what you need: controllable aspect-based opinion summarization of tourist reviews. In: SIGIR '20: proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, pp 1825–1828, https://doi.org/10.1145/3397271.3401269

Mur J, Van Der Plas L (2006) Anaphora resolution for off-line answer extraction using instances. In: Proceedings of the workshop for anaphora resolution (WAR)

Murthy KN, Sobha L, Muthukumari B (2007) Pronominal resolution in tamil using machine learning. In: Proceedings of the first international workshop on anaphora resolution (WAR-I), pp 39–50

Mutso P (2008) Knowledge-poor anaphora resolution system for estonian, Master's thesis, University of Tartu, Tartu, Estonia

Nakaiwa H, Ikehara S (1992) Zero pronoun resolution in a Japanese to English machine translation system by using verbal semantic attributes. In: Proceedings of the third conference on applied natural language processing, association for computational linguistics, pp 201–208

Nastase V, Popescu M (2009) What's in a name? in some languages, grammatical gender. In: Proceedings of the 2009 conference on empirical methods in natural language processing: volume 3, association for computational linguistics, pp 1368–1377

Niraula NB, Rus V, Banjade R, Stefanescu D, Baggett W, Morgan B (2014) The DARE corpus: a resource for anaphora resolution in dialogue based intelligent tutoring systems. In: LREC, pp 3199–3203. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/372_Paper.pdf

Nithya R (2019) Need for anaphoric resolution towards sentiment analysis-a case study with scarlet pimpernel (Novel). Int J Educt Manag Eng 9(1):37–50 (http://www.mecs-press.net), https://doi.org/10.5815/ijeme.2019.01.04

Nøklestad A (2009) A machine learning approach to anaphora resolution including named entity recognition, PP attachment disambiguation, and animacy detection

Novák M (2011) Utilization of anaphora in machine translation. In: Proceedings of contributed papers, Week of Doctoral Students, pp 155–160

Orasan C, St S (2007) Pronominal anaphora resolution for text summarization. In: Proceedings of the recent advances in natural language processing, pp 430–436

Orasan C, Mitkov R, Hasler L (2003) CAST: a computer-aided summarisation tool. In: 10th conference of the European chapter of the association for computational linguistics

Orasan C, Hasler L, St S (2007) Computer-aided summarisation: how much does it really help. In: Proceedings of recent advances in natural language processing (RANLP 2007), pp 437–444

Palomar M, Ferrández A, Moreno L, Martínez-Barco P, Peral J, Saiz-Noeda M, Munoz R (2001) An algorithm for anaphora resolution in Spanish texts. Comput Linguist 27(4):545–567

Palomar M, Civit M, Díaz A, Moreno L, Bisbal E, Aranzabe M, Ageno A, Martí MA, Navarro B (2004) 3LB: construction of a database of syntactic-semantic trees for Catalan, Basque and Spanish. Natural Language Processing, vol 33

Tenney I, Xia P, Chen B, Wang A, Poliak A, McCoy RT, Kim N, Van Durme B, Bowman SR, Das D, Pavlick E (2019) What do you learn from context? Probing for sentence structure in contextualized word representations. In: 7th international conference on learning representations, ICLR 2019, New Orleans, United States

Peng H, Khashabi D, Roth D (2015) Solving hard coreference problems. In: Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies, association for computational linguistics, pp 809–819, https://doi.org/10.3115/v1/n15-1082

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), pp 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers), association for computational linguistics, pp 2227–2237

Phadke M, Devane S (2020) Pronoun resolution task for multilingual machine translation. Available at SSRN 3527387, https://ssrn.com/abstract=3527387 or https://doi.org/10.2139/ssrn.3527387

Poesio M (2004) The MATE/NOME proposals for anaphoric annotation, revisited. In: Proceedings of the 5th SIGdial workshop on discourse and dialogue at HLT-NAACL 2004, pp 154–162

Poesio M, Artstein R (2008) Anaphoric annotation in the ARRAU corpus. In: Proceedings of the sixth international language resources and evaluation (LREC'08). Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/

Poesio M, Kabadjov MA (2004) A general-purpose, off-the-shelf anaphora resolution module: implementation and preliminary evaluation. In: LREC

Poesio M, Ponzetto S, Versley Y (2011) Computational models of anaphora resolution: a survey

Popescu-Belis A, Lalanne D (2004) Reference resolution over a restricted domain: references to documents. In: Proceedings of the ACL-2004 workshop on reference resolution and its applications, association for computational linguistics, Barcelona, pp 71–78

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, pp 1–40

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, association for computational linguistics, pp 777–789

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, pp 2383–2392

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

Recasens M, Martí MA (2010) AnCora-CO: coreferentially annotated corpora for Spanish and Catalan. Lang Resour Eval 44(4):315–345. https://doi.org/10.1007/s10579-009-9108-x

Recasens M, Màrquez L, Sapena E, Martí MA, Taulé M, Hoste V, Versley Y (2010) Semeval-2010 task 1: coreference resolution in multiple languages. In: Proceedings of the 5th international workshop on semantic evaluation, pp 1–8

Reinhart TM (1976) The syntactic domain of anaphora. Doctoral dissertation, Massachusetts Institute of Technology

Reis G, Blair-Goldensohn S, McDonald RT (2014) US Patent No. 8,799,773. Washington: US Patent and Trademark Office

Rich E, LuperFoy S (1988) An architecture for anaphora resolution. In: Proceedings of the second conference on applied natural language processing, pp 18–24

Rodrıguez KJ, Delogu F, Versley Y, Stemle EW, Poesio M (2010) Anaphoric annotation of Wikipedia and Blogs in the live memories corpus. In: Proceedings of LREC, pp 157–163. Retrieved from e:%5CDisser%5CBibliography%5Crodriguez2010anaphoric.pdf

Rusu D, Fortuna B, Grobelnik M, Mladenić D (2009) Semantic graphs derived from triplets pplicwith aation in document summarization. Informatica, 33(3)

Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. In: Advances in neural information processing systems, pp 3856–3866

Saha S, Ekbal A, Uryupina O, Poesio M (2011) Single and multi-objective optimization for feature selection in anaphora resolution. In: Proceedings of 5th international joint conference on natural language processing, pp 93–101

Sakaguchi K, Bras RL, Bhagavatula C, Choi Y (2020) Winogrande: an adversarial winograd schema challenge at scale. In: Proceedings of the thirty-fourth AAAI conference on artificial intelligence 2020, 34(05), pp 8732–8740, https://doi.org/10.1609/aaai.v34i05.6399

Sampson G (2002) English for the computer: the SUSANNE corpus and analytic scheme. MIT Press, Cambridge

Sandhaus E (2008) The new york times annotated corpus. Linguist Data Consort 6(12):e26752

Sandri M, Zuccolotto P (2006) Variable selection using random forests. In: Data analysis, classification and the forward search. Springer, Berlin, pp. 263–270

Savova G, Chapman WW, Zheng J, Crowley RS (2011) Anaphoric relations in the clinical narrative: corpus creation. J Am Med Inform Assoc 18(4):459–465

Segura-Bedmar I, Crespo M, de Pablo C, Martínez P (2009) DrugNerAR: linguistic rule-based anaphora resolver for drug-drug interaction extraction in pharmacological documents. In: Proceedings of the third international workshop on Data and text mining in bioinformatics, ACM, pp 19–26

Senapati A, Garain U (2013) GuiTAR-based pronominal anaphora resolution in Bengali. In: Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short Papers), pp 126–130

Sharaf ABM, Atwell E (2012) QurAna: corpus of the quran annotated with pronominal anaphora. In: LREC, pp 130–137

Shekhar S, Kumar U, Sharma U (2018) To reduce the multidimensionality of feature set for anaphora resolution algorithm. In: Ambient communications and computer systems. Springer, Singapore, pp 437–446, https://doi.org/10.1007/978-981-10-7386-1_38

Sikdar UK, Ekbal A, Saha S, Uryupina O, Poesio M (2015a) Differential evolution-based feature selection technique for anaphora resolution. Soft Comput 19(8):2149–2161. https://doi.org/10.1007/s00500-014-1397-3

Sikdar UK, Ekbal A, Saha S (2015b) Feature selection in anaphora resolution for bengali: a multiobjective approach. In: International conference on intelligent text processing and computational linguistics. Springer, Cham, pp 252–263

Sobha L, Patnaik B (2000) VASISTH: an anaphora resolution system for Indian langauges. In: Proceedings of international conference on artificial and computational intelligence for decision, control and automation in engineering and industrial applications

Stede M, Bieler H, Dipper S, Suriyawongkul A (2006) Summar: combining linguistics and statistics for text summarization. Front Artif Intell Appl 141:827

Steinberger J, Křišťan M (2007) Lsa-based multi-document summarization. In: Proceedings of 8th international workshop on systems and control (vol 7)

Steinberger J, Kabadjov M, Poesio M, Sanchez-Graillet O (2005) Improving LSA-based summarization with anaphora resolution. In: Proceedings of human language technology conference and conference on empirical methods in natural language processing (HLT/EMNLP). The association for computational linguistics, Vancouver, Canada, pp 1–8

Steinberger J, Poesio M, Kabadjov MA, Ježek K (2007) Two uses of anaphora resolution in summarization. Inf Process Manage 43(6):1663–1680

Stojanovski D, Fraser A (2019) Improving anaphora resolution in neural machine translation using curriculum learning. In: Proceedings of machine translation summit XVII Volume 1: Research Track, pp 140–150

Strube M (1998) Never look back: an alternative to centering. ArXiv Preprint Cmp-Lg/9806018, Retrieved from http://arxiv.org/abs/cmp-lg/9806018

Strube M, Müller C (2003) A machine learning approach to pronoun resolution in spoken dialogue. In: Proceedings of the 41st annual meeting on association for computational linguistics-volume 1, association for computational linguistics, pp 168–175

Stuckardt R (2016) Towards a procedure model for developing anaphora processing applications. In: Anaphora resolution, pp 457–484, https://doi.org/10.1007/978-3-662-47909-4_16

Stylianou N, Vlahavas I (2019) A neural entity coreference resolution review. arXiv preprint arXiv:1910.09329

Su J, Yang X, Hong H, Tateisi Y, Tsujii JI (2008) Coreference resolution in biomedical texts: a machine learning approach. In: Dagstuhl seminar proceedings, Schloss Dagstuhl-Leibniz-Zentrum f˙ur Informatik

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

Sun T, Gaut A, Tang S, Huang Y, ElSherief M, Zhao J, Mirza D, Belding E, Chang K-W, Wang WY (2019) Mitigating gender bias in natural language processing: literature review. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 1630–1640

Sundheim BM (1991) Third message understanding evaluation and conference (MUC-3): methodology and test results. In: Natural language processing systems evaluation workshop, pp 1–12

Sundheim BM (1992) Overview of the fourth message understanding evaluation and conference. In: Proceedings of the 4th conference on message understanding, association for computational linguistics, pp 3–21

Tanev H, Mitkov R (2002) Shallow language processing architecture for Bulgarian. In: Proceedings of the 19th international conference on computational linguistics-volume 1, association for computational linguistics, pp 1–7

Tetreault JR (1999) Analysis of syntax-based pronoun resolution methods. In: Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics, pp 602–605, https://doi.org/10.3115/1034678.1034688

Tetreault JR (2001) A corpus-based evaluation of centering and pronoun resolution. Comput Linguist 27(4):507–520. https://doi.org/10.1162/089120101753342644

Ting M, Kadir RA, Azman A, Sembok TMT, Ahmad F (2019) Named entity enrichment based on subject-object anaphora resolution. In: Intelligent computing-proceedings of the computing conference, Springer, Cham, pp 873–884, https://doi.org/10.1007/978-3-030-22868-2_60

Trichelair P, Emami A, Cheung JCK, Trischler A, Suleman K, Diaz F (2018) On the evaluation of commonsense reasoning in natural language understanding. In: Proceedings NeurIPS workshop on critiquing and correcting trends in machine learning, 2018

Tüfekçi P, Kiliçaslan Y (2005) A computational model for resolving pronominal anaphora in turkish using Hobbs' Naïve algorithm. WEC 5:13–17

Uzuner Ö, Solti I, Cadag E (2010) Extracting medication information from clinical text. J Am Med Inform Assoc 17(5):514–518. https://doi.org/10.1136/jamia.2010.003947

Versley Y, Ponzetto SP, Poesio M, Eidelman V, Jern A, Smith J, Moschitti A (2008) BART: a modular toolkit for coreference resolution. In: Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Demo Session, Association for Computational Linguistics, pp 9–12

Vicedo JL, Ferrández A (2000a) Importance of pronominal anaphora resolution in question answering systems. In: Proceedings of the 38th annual meeting on association for computational linguistics, association for computational linguistics, pp 555–562

Vicedo JL, Ferrández A (2000b) Applying anaphora resolution to question answering and information retrieval systems. In: International conference on web-age information management, Springer, Berlin, pp 344–355

Vilain M, Burger J, Aberdeen J, Connolly D, Hirschman L (1995) A model-theoretic coreference scoring scheme. In: Proceedings of the 6th conference on message understanding, Association for Computational Linguistics, pp 45–52

Vlachos A, Gasperin C (2006) Bootstrapping and evaluating named entity recognition in the biomedical domain. In: BioNLP'06: Proceedings of the workshop on linking natural language processing and biology, association for computational linguistics, Morristown, pp 138–145

Vlachos A, Gasperin C, Lewin I, Briscoe T (2006) Bootstrapping the recognition and anaphoric linking of named entities in drosophila articles. In: Proceedings of the Pacific symposium on biocomputing, Hawaii, pp 100–111

Vodolazova T, Lloret E, Muñoz R, Palomar M (2013) Extractive text summarization: can we use the same techniques for any text? In: International conference on application of natural language to information systems. Springer, Berlin, pp 164–175

Voita E, Serdyukov P, Sennrich R, Titov I (2018). Context-aware neural machine translation learns anaphora resolution. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers), Association for Computational Linguistics, pp 1264–1274, https://doi.org/10.18653/v1/p18-1117

Walker MA (1989) Evaluating discourse processing algorithms. In: Proceedings of the 27th annual meeting on association for computational linguistics, pp 251–261, https://doi.org/10.3115/981623.981654

Walker JP (1998) Centering theory in discourse. Oxford University Press, Oxford

Wang Z (2019) MSnet: a BERT-based network for gendered pronoun resolution. In: Proceedings of the first workshop on gender bias in natural language processing, association for computational linguistics, pp 89–95, https://doi.org/10.18653/v1/W19-3813

Wang N, Yuan C, Wong KF, Li W (2002) Anaphora resolution in Chinese financial news for information extraction. In: Proceedings of the 4th world congress on intelligent control and automation, (Cat. No. 02EX527) (vol 3), IEEE, pp 2422–2426

Wang Y, Melton GB, Pakhomov S (2011) It's about "this" and "that": a description of anaphoric expressions in clinical text. In: Proceedings of the American medical informatics association annual symposium (AMIA 2011), Washington, DC, pp 1471–1480

Wang Y, Huang M, Zhu X, Zhao L (2016) Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 606–615

Watson R, Preiss J, Briscoe T (2003) The contribution of domain-independent robust pronominal anaphora resolution to open-domain question-answering. In: Proceedings of the symposium on reference resolution and its applications to question answering and summarization. Venice, Italy, pp 23–25

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

Weston J, Chopra S, Bordes A (2015) Memory networks. In: International conference on learning representations (ICLR)

Wohiduzzaman K, Ismail S (2018) Recommendation system for bangla news article with anaphora resolution. In: 2018 4th International Conference on electrical engineering and information & communication technology (ICEEiCT), pp 467–472, https://doi.org/10.1109/CEEICT.2018.8628075

Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, et al (2016) Google' s neural machine translation system: bridging the gap between human and machine translation. ArXiv Preprint: ArXiv:1609.08144

Wunsch H (2006) Anaphora resolution–What helps in German? In: Pre-proceedings of the international conference on linguistic evidence, Tübingen, Germany, pp 2–4

Xu Y, Yang J (2019) Look again at the syntax: relational graph convolutional network for gendered ambiguous pronoun resolution. In: Proceedings of the first workshop on gender bias in natural language processing, Florence, Italy, Association for Computational Linguistics, pp 96–101, https://doi.org/10.18653/v1/w19-3814

Xu J, Gan Z, Cheng Y, Liu J (2020) Discourse-aware neural extractive text summarization. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 5021–5031

Yadav DS, Dutta K, Singh P, Chandel P (2016) Anaphora resolution for indian languages: the state of the art. Recent Innovations in Science and Engineering 1(2):01–07

Yang X, Su J, Tan CL (2006) Kernel-based pronoun resolution with structured syntactic knowledge. In: Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics, pp 41–48

Yang Q, Yu L, Tian S, Song J (2020) Multi-attention-based capsule network for Uyghur personal pronouns resolution. IEEE Access 8:76832–76840

Yin Q, Zhang W, Zhang Y, Liu T (2017a) A deep neural network for Chinese zero pronoun resolution. In: Proceedings of the 26th international joint conference on artificial intelligence (IJCAI-17), pp 3322–3328, https://doi.org/10.24963/ijcai.2017/464

Yin Q, Zhang Y, Zhang W, Liu T (2017b) Chinese Zero pronoun resolution with deep memory network. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 1309–1318, https://doi.org/10.18653/v1/D17-1135

Yin Q, Zhang Y, Zhang W, Liu T, Wang WY (2018a) Deep reinforcement learning for chinese zero pronoun resolution. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), Melbourne, Australia, Association for Computational Linguistics, pp 569–578, https://doi.org/10.18653/v1/p18-1053

Yin Q, Zhang Y, Zhang W, Liu T, Wang WY (2018b) Zero Pronoun resolution with attention-based neural network. In: Proceedings of the 27th international conference on computational linguistics, pp 13–23

Zeldes A (2017) The GUM corpus: creating multilayer resources in the classroom. Lang Resour Eval 51(3):581–612. https://doi.org/10.1007/s10579-016-9343-x

Zeng J, Ma X, Zhou K (2019) Enhancing attention-based LSTM with position context for aspect-level sentiment classification. IEEE Access 7:20462–20471

Zhang H, Song Y, Song Y (2019a) Incorporating context and external knowledge for pronoun coreference resolution. 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), Association for Computational Linguistics, pp 872–881, https://doi.org/10.18653/v1/N19-1093

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

Zhao S, Ng HT (2007) Identification and resolution of Chinese zero pronouns: a machine learning approach. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)

Zhao J, Wang T, Yatskar M, Ordonez V, Chang K-W (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, Louisiana, Association for Computational Linguistics, pp 15–20, https://doi.org/10.18653/v1/n18-2003

Zhao J, Wang T, Yatskar M, Cotterell R, Ordonez V, Chang KW (2019a) Gender bias in contextualized word embeddings. 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, Minnesota, Association for Computational Linguistics, pp 629–634, https://doi.org/10.18653/v1/n19-1064

Zhao W, Peng H, Eger S, Cambria E, Yang M (2019b). Towards scalable and reliable capsule networks for challenging NLP applications. In Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy, Association for Computational Linguistics, pp 1549–1559, https://doi.org/10.18653/v1/p19-1150

Zheng J, Chapman WW, Crowley RS, Savova GK (2011) Coreference resolution: a review of general methodologies and applications in the clinical domain. J Biomed Inf 44(6):1113-1122, http://www.sciencedirect.com/science/article/pii/S153204641100133X

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B 67(2):301–320

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