Cataphora detection and resolution: Advancements and Challenges in Natural Language Processing

Nabil Moncef Boukhatem, Davide Buscaldi, Leo Liberti

Cataphora detection and resolution: Advancements and Challenges in Natural Language Processing

Nabil Moncef Boukhatem[1 3], Davide Buscaldi[2], and Leo Liberti[3]

[1] OneTeam, Paris, France nboukhatem@oneteam.fr
[2] LIPN CNRS, Université de Paris-Nord, Villetaneuse, France, buscaldi@lipn.univ-paris13.fr
[3] LIX CNRS Ecole Polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau, France, liberti@lix.polytechnique.fr

Abstract. In the field of natural language processing (NLP), accurately understanding and processing complex linguistic structures remains a major challenge. This paper addresses the less-explored phenomenon of cataphora—where a pronoun or noun phrase points forward to a yet-to-be-mentioned entity in the discourse. While anaphora resolution has been extensively studied, cataphora detection and resolution have not received the same level of attention and remain underexplored. This paper seeks to bridge this gap by evaluating state-of-the-art techniques and identifying the obstacles that hinder effective cataphora resolution. We investigate the role of syntactic and semantic ambiguities, contextual influences, and the integration of world knowledge. Additionally, the potential of deep learning, neural network and hybrid models to advance cataphora resolution is explored.

1 Introduction

In the field of natural language processing (NLP), the task of accurately understanding and processing complex linguistic structures [36] remains a major challenge. Among these complex structures, anaphora [46] and cataphora [2, 49] are two linguistic phenomena that play a critical role in determining the coherence and meaning of a text. While anaphora [30] refers to the use of a pronoun or noun phrase that points back to a previously mentioned entity, cataphora is the reverse: it is the reference of a pronoun or noun phrase that points forward to a yet-to-be-mentioned entity in the discourse. Although anaphora resolution [35] has been extensively studied in NLP, cataphora detection and resolution have not received the same level of attention [2]. This research paper aims to bridge this gap by exploring state-of-the-art techniques for cataphora detection and resolution, as well as highlighting the current challenges and future directions in this area.

In this paper, we will first provide an overview of the existing methods and algorithms for cataphora detection and resolution, discussing their strengths and

weaknesses. Next, we will delve into the challenges and limitations faced by current approaches, with a particular focus on the impact of syntactic and semantic ambiguities, as well as the influence of context and world knowledge on cataphora resolution. We will also examine the role of deep learning and neural network models in advancing the field, exploring their potential for addressing the existing challenges in cataphora resolution. Finally, we will outline possible directions for future research, emphasizing the need for more comprehensive evaluation metrics, larger and more diverse annotated corpora, and novel techniques that can better capture the complex interplay between linguistic and contextual cues in cataphora resolution.

2 Background and Related Work

2.1 Definition and Characteristics of Cataphora

Cataphora [49] is a linguistic phenomenon in which a pronoun or a noun phrase (NP) refers to an entity that has not yet been introduced and that appears later in the text or discourse. It creates a dependency that must be resolved to fully comprehend the intended meaning and that can be subject to syntactic and semantic constraints, including agreement in number, gender, and person, as well as syntactic and semantic compatibility between the cataphoric expression and its antecedent. This forward-looking reference is the opposite of anaphora, where a pronoun or NP points back to a previously mentioned entity. Cataphora plays an essential role in establishing coherence and meaning within a text, as it facilitates the identification of entities and their relationships. Cataphoric references [2] can be found in various syntactic structures and can involve different types of expressions, including pronouns, demonstratives, and definite NPs. The most common form of cataphora involves pronouns, such as *"he,"* *"she,"* *"it,"* or *"they,"* which refer to a subsequent noun phrase. For example, in the following sentence, the pronoun *"he"* is a cataphoric reference to *"John."*:

"Before *he* left the house, *John* grabbed his umbrella,"

The sentence is an example of strict cataphora, in that it makes use of a pronoun to refer to an antecedent. Non-strict cataphora apply the same structure, but with a noun or noun phrase as the cataphora instead of a pronoun.

Cataphoric references often occur within specific discourse structures [6], such as topic introduction, summary statements, or contrastive elements. Accurately resolving cataphoric expressions is challenging because it relies heavily on contextual information and world knowledge that NLP systems may not have access to. Despite this, identifying and resolving cataphoric expressions is essential for various NLP applications, as it helps represent the semantic relationships between entities in a text.

2.2 Cataphora types

Linguistic terminology and concepts can be nuanced, with variations in how terms like cataphora are applied or understood in different contexts, languages or among different scholars.

In the English language, cataphora manifests in several distinct forms. Understanding these types and shedding light on their linguistic properties, syntactic structures, and computational implications is essential for our experiments.

Pronominal Cataphora Pronominal cataphora occurs when a pronoun precedes and refers to a noun or noun phrase later in a sentence or discourse. Example: "Even if *she* doesn't admit it, *Mary* will always know the truth."

Computational approaches to pronominal cataphora entail sophisticated algorithms for coreference resolution, which aim to identify and link pronouns with their corresponding antecedents in text corpora.

Advanced NLP models, especially those based on deep learning and trained on large corpora, can identify and link "she" to "Mary" by understanding sentence structure and context.

Determiner Cataphora This type involves a determiner (e.g., this, that, these, those) pointing to a noun phrase that appears later.

Example: "In *that* scenario, the team would need to score in the final minutes to win."

Similar to pronominal cataphora, determiner cataphora can be addressed through coreference resolution techniques. The models need to understand the context in which the determiner is used and link it to the appropriate noun phrase, a task that might require more contextual information and understanding of the discourse structure.

Lexical Cataphora Lexical cataphora occurs when a word or phrase refers to a more specific term or explanation that follows.

Example: "*The punishment*, *three days of suspension*, was harsh but fair."

Addressing lexical cataphora involves recognizing the introductory phrase and linking it to its explanatory or specifying clause. This can be challenging and may require syntactic parsing to identify the structure of sentences and semantic analysis to understand the relationship between the general term and its specific explanation.

Clause Cataphora Clause cataphora occurs when a dependent clause precedes its main clause within a sentence or discourse. This type of cataphora often provides a reason, condition, or outcome that is explained or concluded subsequently, facilitating discourse coherence and narrative progression.

Example: "*To secure victory*, while facing sush a harsh competition, *the athlete must outperform her last best performance.*"

Computational approaches to clause cataphora necessitate sophisticated parsing algorithms capable of identifying and analyzing syntactic dependencies between clauses in complex sentences.

Advanced NLP techniques, such as dependency parsing combined with semantic role labeling, can help in understanding the causal or conditional relationships between the clauses.

Nominal Cataphora Nominal cataphora involves the use of a noun phrase or noun that precedes its referent, pointing forward to a more detailed explanation or specification. This type of cataphora is often employed to introduce descriptive noun phrases, followed by the explicit mention of the referent later in the discourse. This type can overlap with pronominal and determiner cataphora but is distinct in its focus on nouns or noun phrases.

Example: "*This idea*, that *you can achieve anything through hard work*, is widely promoted in society."

Computational models for nominal cataphora necessitate robust techniques for noun phrase identification and coreference resolution, particularly in tasks requiring fine-grained semantic analysis, to identify the introductory nominal phrase ("This idea") and its subsequent detailed explanation ("that you can achieve anything through hard work").

This may require not just syntactic understanding but also semantic processing to grasp the nature of the explanation or specification.

Adverbial cataphora Adverbial cataphora manifests when an adverbial phrase precedes the action, state or event it modifies, creating anticipation for the subsequent description of the action. This type of cataphora plays a crucial role in structuring discourse temporally, guiding the reader or listener's interpretation of narrative events, as it often sets the stage for how, when, where, or why something happens.

Example: "*To make matters worse*, after all the challenges he faced, *the main speaker canceled at the last minute.*"

Handling adverbial cataphora effectively requires understanding the temporal, locational, causal, or manner information provided by the adverbial phrase and how it relates to the subsequent action or state. Dependency parsing, to identify the relationship between sentence components, and semantic role labeling, to understand the contextual roles of these components, are crucial. The goal is to link the adverbial phrase ("To make matters worse") with the event it describes ("the main speaker canceled at the last minute").

2.3 Importance of Cataphora Resolution in NLP Applications

Accurately identifying and resolving cataphoric references is an essential step in NLU and has significant implications for various NLP applications, as it contributes to the accurate representation of the semantic relationships between entities in a text. These applications include machine translation, information

extraction, text summarization, dialogue systems, and coreference resolution for example.

In machine translation, the resolution of cataphoric expressions plays a pivotal role in maintaining coherence and preserving the intended meaning of a text during translation into another language. Similarly, in information extraction systems, it helps establish connections between entities and extract more accurate and complete information from unstructured text. This process is also crucial for generating concise and coherent summaries in text summarization applications and essential for dialogue systems reliant on NLU to interpret user input and generate appropriate responses. Lastly, resolving cataphoric expressions is an essential part of coreference resolution [42, 5], which is the task of identifying and linking various expressions referring to the same entity within a text.

Overall, accurately resolving cataphoric expressions is essential for NLP systems to better understand and process textual information, leading to improved performance and more sophisticated NLU capabilities.

2.4 Overview of Anaphora Resolution and its Relation to Cataphora

Although anaphora and cataphora have differences, they share commonalities that make their resolution tasks interconnected and complementary in NLP applications. Both require syntactic and semantic constraints to determine the compatibility between the referring expression and its antecedent. They also contribute to establishing coherence and maintaining the discourse structure within a text, identifying entities and their relationships. Additionally, both tasks pose similar challenges, requiring contextual information and world knowledge, which actual NLP systems may have limited access to, even the most advanced ones.

Anaphora resolution has historically received more attention in NLP research than cataphora's due to the higher frequency of anaphoric expressions, their relevance to various NLP applications, the availability of data, and the historical focus on anaphora in linguistics. This resulted in the development of various approaches such as rule-based, statistical, and machine learning-based methods [52, 36]. These techniques have significantly contributed to improving coreference resolution [42] systems and integrating them into NLP applications.

In contrast, cataphora resolution has not been extensively studied, and therefore, lags behind in terms of methodological advancements. However, researchers can use the progress made about anaphora's as a valuable foundation and by building on existing knowledge and techniques developed, they can advance cataphora resolution and develop accurate and robust methods. This advancement will further enhance the performance of existing coreference resolution and natural language understanding systems.

3 Existing Approaches to Cataphora Detection and Resolution

3.1 Rule-based Methods

Rule-based methods for cataphora detection and resolution use predefined linguistic rules and heuristics to identify and link cataphoric expressions to their antecedents. These methods rely on linguistic theories and expert knowledge to provide insights into the syntactic, semantic, and discourse-level constraints governing cataphoric references [23]. While rule-based approaches have been widely used in early NLP research and can be highly accurate, they have limitations.

Syntactic rules are often used in rule-based methods to identify potential antecedents for cataphoric expressions by taking into account the syntactic structure of sentences, the grammatical relations between constituents [21], and the linear order of elements in the text. Morphological agreement is also incorporated to require agreement between the cataphoric expression and its antecedent in terms of number, gender, and person [21, 1]. To further refine the list of candidate antecedents, semantic constraints may be applied based on the compatibility of the cataphoric expression and its potential antecedent. These methods may take into account the discourse structure and coherence of the text to resolve cataphoric references [25].

Despite their effectiveness in specific contexts and languages, rule-based methods have limitations [39]. They lack flexibility, may not generalize well to different languages, genres, or domains, and require expert knowledge and manual adjustments. They are sensitive to errors in the input data and can be computationally expensive, especially when dealing with large-scale data or complex linguistic phenomena. Nevertheless, they provide a foundation for the development of more advanced techniques for cataphora resolution, such as statistical and machine learning approaches [23, 26].

3.2 Statistical and Machine Learning Techniques

Researchers have turned to statistical and machine learning (ML) techniques to address the limitations of rule-based methods [23, 26, 24]. These approaches are characterized by their ability to automatically learn patterns and features associated with cataphoric references from annotated training data.

Feature extraction is a critical step in statistical and ML methods, and relevant features can include syntactic, semantic, and discourse-related properties like syntactic patterns, positional information of pronouns, semantic cues, the distance between the pronoun and the candidate antecedents [43]. Supervised learning techniques, such as Support Vector Machines, Decision Trees, and Naive Bayes classifiers, are commonly used in cataphora resolution, but unsupervised and semi-supervised learning techniques can also be employed [12]. Evaluation metrics such as precision, recall, F1-score, and accuracy are used to assess the performance of these methods.

While statistical and ML techniques offer several advantages over rule-based methods, such as their ability to learn from data and handle variations in language, genre, or domain, they also come with their own set of challenges and limitations [12]. These include their dependence on annotated data, the challenge of identifying relevant and informative features, and the lack of interpretability and explainability in the resulting models. Despite these challenges, statistical and machine learning techniques have contributed significantly to the advancement of cataphora resolution and paved the way for more advanced approaches, such as neural network models and deep learning.

3.3 Neural Network Models and Deep Learning Approaches

Neural network models have significantly improved various NLP tasks, which includes cataphora detection and resolution, by automatically learning feature representations and capturing complex patterns in the data. These models offer better performance, scalability, and adaptability compared to ML methods after being trained on large datasets and learnt to link pronouns with their referents, whether they are anaphoric or cataphoric in nature.

Word embeddings and contextualized representations are used to represent cataphoric expressions and their potential antecedents, capturing the semantic and syntactic properties of the text [41]. Sequence models, such as RNNs and LSTMs, model the dependencies and relationships between entities in a sequence, while attention mechanisms enhance these models by selectively focusing on relevant parts of the input.

End-to-end learning [27] and transfer learning techniques, such as fine-tuning pre-trained language models, have further improved the performance and generalizability of neural network models for cataphora resolution and shown promising results [17].

However, these approaches also face challenges, such as the need for significant computational resources and difficulties in interpretability and explainability. Furthermore, neural network models may struggle to capture fine-grained linguistic knowledge and specific linguistic constraints that govern cataphora resolution.

In summary, deep learning techniques have shown significant potential in advancing the research field around cataphoras. Future research will aim to address the challenges and limitations associated with these approaches, with the goal of developing more accurate, efficient, and interpretable models.

4 Challenges and Limitations in Cataphora Detection and Resolution

4.1 Syntactic and Semantic Ambiguities

Cataphora resolution, as a subtask of coreference resolution, is inherently affected by lexical, syntactic and semantic ambiguities in language[29, 3], who's

presence is one of its primary challenges. These ambiguities pose challenges for both human readers and computational models attempting to identify and link cataphoric expressions to their antecedents.

Cataphoric expressions are often complex and ambiguous, making it difficult for computational models to accurately identify and resolve them. For instance, the pronoun *"He"* in the sentence "When *He* started bleeding, *Michael* was starring at *Pierre* through the open window." could refer to either to *Micheal* or to *Pierre*, depending on the context. Similarly, in the sentence "John tried fixing *it*, but *the printer* still prints *the ticket* wrong.", the pronoun *"it"* could refer to either *the printer* or *the ticket*. Understanding and resolving these types of ambiguities is essential for accurate NLU in NLP applications. They can arise from various sources, such as structural and coordination ambiguities, lexical and predicate ambiguities, and cataphoric ambiguity.

Structural ambiguity refers to cases where a sentence can be parsed in multiple ways, leading to different syntactic structures [3]. This ambiguity can create uncertainty regarding the grammatical role of a noun phrase, making it difficult to determine whether it serves as a valid antecedent for a cataphoric expression. Coordination ambiguity occurs when it is unclear how conjuncts in a coordinated structure should be grouped or related. This ambiguity can create uncertainty about the scope and relationships of potential antecedents within a coordinated structure.

Lexical ambiguity arises when a word or phrase has multiple meanings or senses [29]. This ambiguity can affect cataphora resolution, as the correct interpretation of the ambiguous word or phrase is crucial for determining the semantic compatibility between a cataphoric expression and its potential antecedent. Predicate ambiguity occurs when it is unclear how the arguments of a predicate should be assigned or interpreted. This ambiguity can impact cataphora resolution, as the relationships between cataphoric expressions and their potential antecedents may depend on the correct interpretation of the predicate and its arguments.

Cataphoric ambiguity arises when a cataphoric expression can be linked to multiple potential antecedents. This ambiguity can make it challenging to determine the correct antecedent for a cataphoric expression, as the models need to consider various syntactic, semantic, and discourse-level factors to resolve the ambiguity. Overall, resolving these ambiguities requires advanced natural language understanding models that can handle the complexity and variability of natural language text.

4.2 Contextual and World Knowledge Constraints

Cataphora resolution is not only influenced by syntactic and semantic ambiguities [3] but also by contextual and world knowledge constraints [31, 20]. These constraints are crucial for human language comprehension, aiding in disambiguating potential antecedents and establishing coherent and meaningful interpretations of text. Incorporating these constraints in cataphora resolution

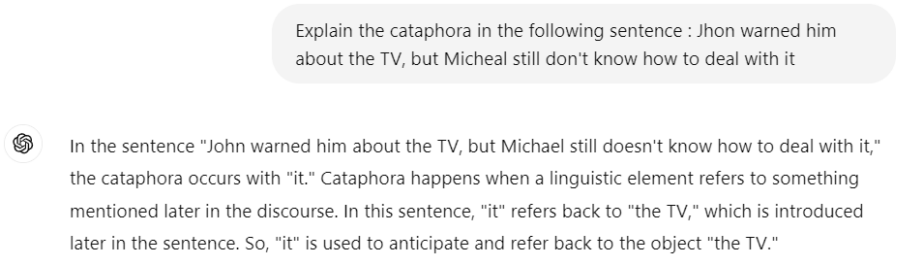


Fig. 1: An exemple of GPT 3.5 wrongly identifying and resolving an anaphora instead of a cataphora

models is essential for developing accurate and robust systems capable of understanding and processing natural language in a manner that resembles human comprehension.

Contextual constraints can be classified into local and global contexts. The immediate context surrounding a cataphoric expression can provide valuable information for identifying and linking the expression to its correct antecedent, such as the grammatical role of the referring expression and the coherence relations between clauses or sentences. The broader context of a text or discourse can also provide cues for resolving cataphoric expressions, such as the topic, theme, or genre of the text.

World knowledge constraints can be classified into commonsense knowledge and domain-specific knowledge. Commonsense knowledge refers to the general knowledge about the world that humans typically possess, such as facts about objects, properties, events, and relationships. Domain-specific knowledge pertains to specialized knowledge about particular fields, such as medicine, law, or sports. Incorporating this knowledge in cataphora resolution models can help to improve their accuracy and robustness, particularly when processing texts from specific domains.

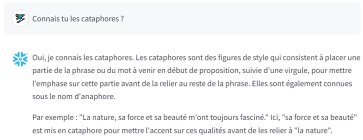
To integrate these constraints into cataphora resolution models, researchers can leverage knowledge bases, ontologies, and commonsense reasoning resources like ConceptNet [44] or Cyc [28]. They can also use advanced computational models that can automatically learn to capture and exploit contextual and world knowledge from the data. Addressing these challenges will enable researchers to continue advancing the state of the art, leading to more sophisticated and human-like NLU systems.

4.3 Cross-linguistic, Multilingual and Genre-specific Variations

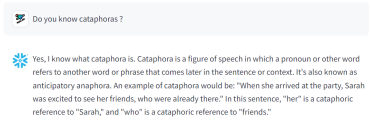
Cross-linguistic and Multilingual Variations in Cataphoric References Cross-linguistic and multilingual variations underscore a significant challenge in the field of NLP. The variability in the use and interpretation of cataphoric references across languages can be attributed to distinct syntactic structures, pronoun systems, and linguistic conventions [1, 39, 41, 8, 19, 11]. For instance,

whereas English may exhibit stringent syntactic rules that confine cataphora to specific contexts, other languages might afford a more lenient syntax that accommodates a wider utilization of cataphoric references. Furthermore, the semantic intricacies of pronouns in various languages necessitate models that possess an acute understanding of these nuances to accurately resolve references.

Addressing this challenge necessitates the utilization of comprehensive linguistic datasets that encapsulate the syntactic and semantic diversity across languages. These datasets should facilitate a deep understanding of pronoun placement, referent relationships, and how syntactic structures influence interpretation. Additionally, the semantic properties of pronouns and referring expressions across different languages must be meticulously accounted for. This entails interpreting not only the syntax but also the semantic cues that are pivotal for the effective resolution of cataphoric references [1, 39, 41, 8, 19, 11, 4].



(a) Arctic LLM giving the wrong definition of a cataphora in French



(b) Arctic LLM giving the correct definition of a cataphora in English

Fig. 2: The impact of language on LLMs behavior

Genre-specific Variations and Their Implications While genre-specific variations indeed influence the use and interpretation of cataphora, they accentuate the need for computational models to be adaptable to the contextual nuances of various texts—ranging from literary compositions to conversational transcripts. However, the impact of these variations is relatively low compared to other challenges.

Advancing Cataphora Resolution Systems: A Multifaceted Approach The advancement of cataphora resolution systems that are adept in navigating both cross-linguistic complexities and genre-specific nuances requires a comprehensive and multifaceted approach. This encompasses the creation of diverse, linguistically rich training datasets and the development of models capable of dynamically adjusting their interpretative strategies based on the linguistic and stylistic context of a given text. Employing techniques such as transfer learning—where a model trained on one language or genre is adeptly adapted to another—and multi-task learning, which facilitates learning from related tasks such as genre classification or language identification, stands as a beacon of progress in enhancing the capabilities of cataphora resolution systems [55, 13, 33, 11, 53].

In conclusion, the pursuit of developing robust and accurate cataphora resolution systems necessitates an integrative approach that accounts for the vast cross-linguistic and multilingual variations.

5 Evaluation Metrics and Benchmark Datasets

5.1 Overview of Evaluation Metrics for Cataphora Resolution

Evaluating the performance of cataphora resolution models is crucial for identifying the strengths and weaknesses of different approaches and for guiding the development of more accurate and robust methods. In this section, we provide an overview of the main evaluation metrics used to assess the performance of cataphora resolution techniques.

a. Precision: Precision measures the proportion of correctly resolved cataphoric expressions among all the expressions identified by the model as cataphora. A high precision indicates that the model has a low rate of false positives, i.e., it rarely misidentifies non-cataphoric expressions as cataphora or links cataphoric expressions to incorrect antecedents.

b. Recall: Recall measures the proportion of true cataphoric expressions that were successfully resolved by the model. A high recall indicates that the model has a low rate of false negatives, i.e., it rarely misses cataphoric expressions or fails to resolve them.

c. F1-score: F1-score [10] is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between these two measures. A high F1-score indicates that the model performs well in both precision and recall, achieving a good balance between minimizing false positives and false negatives.

d. Accuracy: Accuracy is the proportion of correct predictions made by the model, considering both true positives (correctly resolved cataphora) and true negatives (correctly identified non-cataphoric expressions). A high accuracy indicates that the model performs well overall in identifying and resolving cataphoric expressions.

5.2 Evaluation Datasets

Several benchmark datasets have been developed to facilitate the evaluation and comparison of anaphora resolution methods. These datasets typically consist of annotated texts, where anaphoric expressions and their corresponding antecedents are labeled. Examples of such datasets include the OntoNotes corpus, or the Anaphora Resolution EXercise (AREx) corpus. Although these datasets primarily focus on anaphora resolution, they can also be used to evaluate cataphora resolution techniques by adapting the annotations and evaluation protocols accordingly.

For our experiment, we structured and will rely on a handcrafted dataset composed of three parts : one subset of 300 sentences with a cataphoric reference, one subset of 200 with an anaphoric reference and another subset of 100 without any reference.

6 Comparative Analysis

6.1 Compared methods

We have tried exploring various methodologies to address the inherent challenges posed by this linguistic phenomenon. Among these, three main approaches stand out: coreference resolution models [42, 5], pre-trained large language models (LLMs) [34], and hybrid methods that we propose that are a combination of rule-based systems with the aforementioned models. Each method presents its own set of advantages and limitations, with mixed results.

Coreference models Coreference resolution models [42, 5] represent a foundational approach in NLP. These models are engineered to identify and reconcile referential expressions within texts, employing syntactic and semantic analysis to establish the relationships between pronouns and their antecedents. The efficiency of these models is largely attributable to their algorithmic precision in parsing and understanding structured linguistic data. However, their effectiveness can be constrained by the linguistic diversity and complexity of the texts under analysis, highlighting a potential limitation in their adaptability to varied linguistic contexts. The academic discourse surrounding these models often centers on their capability to generalize across languages and genres, an area ripe for further exploration and development.

The models explored cover the Stanza coreference resolution module (CRM) [32], AllenNLP CRM [9] and FastCoref [38], that offer specialized capabilities in identifying and linking referential expressions within text. These models leverage syntactic and semantic analysis to discern the relationships between pronouns and their antecedents, providing a foundation for cataphora resolution. However, their performance can vary based on the linguistic characteristics and complexity of the texts they analyze.

Pre-trained LLMs The advent of pre-trained LLMs has marked a paradigm shift in NLP. These models, known for their capacity for zero-shot and few-shot learning [22, 18] and potentially augmented through fine-tuning offer a comprehensive understanding of language structure and context [34]. The principal academic interest in LLMs lies in their ability to grasp and apply nuanced language patterns without explicit programming, a testament to their advanced algorithmic underpinnings. This adaptability makes LLMs particularly effective in complex linguistic scenarios, surpassing traditional models in their contextual awareness and flexibility. Nonetheless, the computational resource intensity and the need for extensive data during the training phase remain significant considerations in their deployment, also considering fine-tuning perspectives, while few-shot learning can be seen as a viable alternative.

GPT-3.5, GPT-4 [37], GPT-4o, Mistral Large [15, 16], and Llama3 [47, 48], represent cutting-edge and the most performant LLMs, as they bring a vast understanding of language structure and context to bear on the task of cataphora resolution and have been selected for our experiments.

Hybrid methods Hybrid methods present an innovative approach by combining the precision of rule-based systems with the contextual understanding of either coreference models or LLMs. This methodology initiates with a rule-based layer that identifies potential antecedents, such as pronouns, nouns, and clauses, integrating linguistic rules for gender and number agreement to refine the focus of analysis. This preparatory phase is critical in optimizing the subsequent computational process, directing the model's resources towards the most probable candidates, significantly reducing the incidence of false positives and enhancing overall accuracy. This approach should showcase a promising avenue for enhancing the accuracy and efficiency of cataphora resolution. This balanced approach underscores the potential for synergistic methodologies in addressing the complexities of the linguistic phenomena.

6.2 Experiment methodology

In our study, the construction of our dataset involved three distinct phases. Initially, we manually annotated references, focusing on identifying both anaphoric and cataphoric elements within the sentences.

For each subset, we generated a set of candidate pronouns and references by applying rule-based methods. These rules were designed to identify various linguistic elements, including possessive pronouns, personal pronouns, common nouns, proper nouns, and nominal groups. By leveraging these rules, we established a comprehensive pool of candidate references to facilitate subsequent analysis.

Following the generation of candidate sets, we proceeded to create two other distinct subsets. The first subset comprised 100 examples from each dataset segment, arranged sequentially (i.e., 100 anaphoric examples followed by 100 cataphoric examples and then 100 sentences without references), for a total of 300 examples. The second subset encompassed all examples arranged in a random order. This randomization aimed to assess the impact of contextual variation on the performance of LLMs, thereby ensuring comprehensive evaluation and analysis of the model's capabilities in handling diverse linguistic contexts.

To evaluate the performance of the selected models, we devised the following experimental setup:

Firstly, three state-of-the-art neural coreference models, namely FastCoref, Stanza, and AllenNLP, were applied to the three original subsets of our dataset. Since the results of these models are unaffected by the distribution of sentences, they were initially employed for preliminary assessment. Among these models, Stanza emerged as the most effective coreference model. The sentences were later provided along with candidates obtained using rule-based methods to ease the identification of cataphorical references.

Secondly, we assessed the performance of selected LLMs, including GPT 3.5, GPT 4, GPT 4o, Llama3 and Mistral Large, under two distinct scenarios: 1-shot and 5-shot situations. Initially, these models were evaluated on the three separated datasets to gauge their proficiency in cataphora resolution. Subsequently,

the evaluation was extended to the merged dataset and the dataset with random distribution to ascertain the models' robustness across varied contexts.

The evaluation prompt provided to these models was structured to elicit their ability to identify cataphoric references within sentences. Specifically, the prompt directed the models to list cataphoric references, separating the original sentence and the identified cataphoric references by a semicolon. Additionally, the prompt accommodated the possibility of multiple cataphoric references within a single phrase, necessitating the listing of such references in a structured format.

An example provided within the prompt clarified the expected response format, demonstrating how the original sentence and the corresponding cataphoric references should be delineated. This standardized prompt ensured consistency in the evaluation process across all LLMs tested.

The 1-shot prompt used took this form :

"I will give you a list of sentences that might contain cataphoric references, I want you to identify these cataphoric references and list them, separating the original sentence and the cataphoric references by a ; and there can be multiple cataphoric references in a same phrase, so make it a list of couples separated by commas. I give you an example : He was so embarrassed, Harry was so ashamed;[(He,Harry)]"

Finally, in the hybrid scenario for LLMs, we introduced a modification to the evaluation prompt. We indicated that at the conclusion of each sentence, a proposition of potential pronouns was provided, serving as the sole candidates to be considered if a cataphoric reference existed within the sentence. This modification aimed to constrain the scope of their search to the provided pool only during the resolution process.

All the results for all the different approaches have been manually annotated.

7 Computational results, analysis and discussion

7.1 Neural Coreference Resolution approach

Models	**Standard**				**Hybrid**			
	Precision	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Stanza	**45.45%**	**68.11%**	**54.52%**	26.83%	**50.63%**	**66.45%**	**57.47%**	35.50%
FastCoref	39.70%	44.19%	41.82%	33.67%	38.79%	42.52%	40.57%	33.00%
AllenNLP	42.74%	49.83%	46.01%	**37.33%**	42.57%	49.50%	45.78%	**37.33%**

Table 1: Metrics for Neural Coreference Resolution approach

Stanza Hybrid Stanza shows better precision and F1-Score compared to the standard model, indicating the effectiveness of the hybrid approach. However, the overall success rate remains low, highlighting underlying issues in resolving

complex cataphors. Stanza is particularly sensitive to the order of sentences; changing the order can disrupt the model, leading to hallucinations and incorrect responses. Additionally, Stanza struggles with sentences containing multiple cataphors, making it challenging to handle references correctly. The Mean Reciprocal Rank (MRR) improves with the hybrid model, suggesting better precision in cataphor resolution.

To optimize Stanza, enhancing the model to manage multiple contexts and reduce hallucinations could improve performance. Stabilizing responses regardless of sentence order could help reduce sensitivity and errors.

AllenNLP Performance between the standard and hybrid model of AllenNLP is similar, with the standard model showing a slight edge. This indicates that the hybrid approach does not provide significant benefits for AllenNLP. Therefore, efforts should focus on improving the standard model to enhance precision and recall.

FastCoref In FastCoref, standard model shows slightly better recall and F1-Score compared to hybrid model, with a higher success rate. Thus, the enhancement of the model, particularly in handling complex sentences and avoiding reference errors, is recommended.

7.2 LLM Based Cataphora Resolution approach

Batch Examples The 5-shot approaches in LLMs generally show superior performance in terms of F1-Score and success rate, indicating robustness and effectiveness in resolving cataphors. In contrast, 1-shot and hybrid approaches are less effective, with hybrids sometimes introducing unnecessary references by forcing themselves to identify references that are not existant. Specific observations include GPT-4o in 1-shot, which shows low precision but perfect recall, and Llama3 and Mistral Large, which perform exceptionally well in 5-shot configurations.

Merged Examples When sentences are ordered, LLMs, particularly in 5-shot, demonstrate greater stability. GPT-4 and GPT-4o show a good improvement in 5-shot. Llama3 and Mistral Large are also more robust in 5-shot, showing less sensitivity to sentence order and maintaining high performance even with randomized examples.

Randomized Examples In randomized examples, the loss and generation of phrases often disrupt results. For instance, Mistral Large in 1-shot sometimes repeats phrases without correct resolution. The 5-shot approaches outperform 1-shot and hybrids, which sometimes refer to parts of sentences incorrectly, indicating a need for refinement to avoid non-existent relationships.

Models	Precision	Recall	F1-Score	Accuracy
GPT-3.5 1-shot	71.36%	**99.67%**	**83.17%**	68.67%
GPT-3.5 5-shot	66.23%	**100.00%**	79.68%	63.33%
GPT-3.5 Hybrid	70.62%	**99.33%**	82.55%	59.50%
GPT-4 1-shot	54.82%	83.33%	66.14%	55.17%
GPT-4 5-shot	50.21%	79.00%	61.40%	49.00%
GPT-4 Hybrid	56.90%	88.00%	69.11%	58.33%
GPT-4o 1-shot	0.78%	**100.00%**	1.56%	7.83%
GPT-4o 5-shot	**72.54%**	96.00%	**82.64%**	**78.50%**
GPT-4o Hybrid	64.03%	89.00%	74.48%	64.17%
Llama3 1-shot	71.69%	90.33%	79.94%	**70.50%**
Llama3 5-shot	**75.70%**	**98.67%**	**85.67%**	**71.50%**
Llama3 Hybrid	55.65%	90.33%	68.87%	29.33%
Mistral Large 1-shot	60.12%	**100.00%**	75.09%	16.83%
Mistral Large 5-shot	**77.84%**	**98.33%**	**86.89%**	**82.83%**
Mistral Large Hybrid	58.59%	94.33%	72.29%	43.83%

Table 2: Metrics for LLM-based Cataphora Resolution Methods (Batch Treatment)

Models	Precision	Recall	F1-Score	Accuracy
GPT-3.5 1-shot	71.36%	**99.67%**	**83.17%**	68.67%
GPT-3.5 5-shot	66.23%	**100.00%**	79.68%	63.33%
GPT-3.5 Hybrid	70.62%	**99.33%**	**82.55%**	59.50%
GPT-4 1-shot	54.82%	83.33%	66.14%	55.17%
GPT-4 5-shot	50.21%	79.00%	61.40%	49.00%
GPT-4 Hybrid	56.90%	88.00%	69.11%	58.33%
GPT-4o 1-shot	0.78%	**100.00%**	1.56%	7.83%
GPT-4o 5-shot	**72.54%**	96.00%	**82.64%**	**78.50%**
GPT-4o Hybrid	64.03%	89.00%	74.48%	64.17%
Llama3 1-shot	71.69%	90.33%	79.94%	70.50%
Llama3 5-shot	**75.70%**	**98.67%**	**85.67%**	**71.50%**
Llama3 Hybrid	55.65%	90.33%	68.87%	29.33%
Mistral Large 1-shot	60.12%	**100.00%**	75.09%	16.83%
Mistral Large 5-shot	**77.84%**	98.33%	**86.89%**	**82.83%**
Mistral Large Hybrid	58.59%	94.33%	72.29%	43.83%

Table 4: Metrics for LLM-based Cataphora Resolution approach (Randomized Treatment)

Models	Precision	Recall	F1-Score	Accuracy
GPT-3.5 1-shot	56.27%	**96.96%**	71.22%	51.26%
GPT-3.5 5-shot	71.60%	79.18%	**75.20%**	64.32%
GPT-3.5 Hybrid	52.63%	3.45%	6.47%	50.42%
GPT-4 1-shot	74.67%	75.42%	**75.04%**	**75.04%**
GPT-4 5-shot	**83.96%**	52.86%	64.88%	71.36%
GPT-4 Hybrid	61.03%	**95.96%**	74.61%	63.48%
GPT-4o 1-shot	69.29%	85.86%	**76.69%**	**73.37%**
GPT-4o 5-shot	74.38%	80.13%	**77.15%**	**76.05%**
GPT-4o Hybrid	72.88%	75.08%	73.96%	67.34%
Llama3 1-shot	70.89%	56.76%	63.04%	62.98%
Llama3 5-shot	**84.07%**	51.52%	63.88%	64.49%
Llama3 Hybrid	58.03%	**97.31%**	72.70%	39.87%
Mistral Large 1-shot	52.88%	**95.96%**	68.18%	23.12%
Mistral Large 5-shot	71.64%	82.49%	**76.68%**	**74.04%**
Mistral Large Hybrid	58.27%	**97.31%**	72.89%	56.45%

Table 3: Metrics for LLM-based Cataphora Resolution approach (Merged Treatment)

Other observations In various scenarios, LLMs often encounter difficulties that affect their performance. One notable issue is that LLMs can be misled by the rule-based indications, where they may correctly identify the reference but fail to select the appropriate pronoun. Mistral Large, for example, struggles significantly with nominal phrases and traps, presenting a hard challenge for all models.

When processing random examples, a substantial number of sentences are either lost or new sentences are inadvertently generated, which disrupts the results. Mistral Large 1-shot, in particular, has a tendency to repeat the phrase without progressing towards a correct resolution. Moreover, hybrid models sometimes incorrectly reference parts of the reference instead of the whole, constructing non-existent references. For instance, in the sentence "Without considering its impact, the decision was hastily made," the hybrid model might incorrectly reference "decision's impact" rather than correctly pointing to "decision."

To improve performance, it's essential to incorporate decomposition rules based on punctuation in addition to managing parsing, syntax, and semantics.

Some models, given a long input, are unable to respond correctly and exhibit altered behaviors. For example, Llama3 becomes highly erratic when using hybrid models.

Initial tests in French have shown promising directions for future improvements, with varying processing speeds across different models and offering more

tools, due to French language specifics, to identify the candidates and making them match with the references, especially genre and number.

Consider the following processing speed observations: Meta can finish inference first if it avoids generating additional irrelevant sentences, but it might fail last due to the window size limitation. GPT-4o, with a large processing window, typically performs robustly, whereas GPT-3.5, despite having a smaller window, operates faster. Mistral Large, with the largest window, takes longer to process input, and excessive input significantly impacts its behavior. This pattern is consistent across models, where maintaining a maximum input length of 100 sentences, sometimes 50, is crucial for coherence. This constraint is particularly relevant in random hybrid approaches.

Lastly, attempts to enhance performance through embedding-based methods combined with rule-based techniques have not yielded satisfactory results, indicating that further refinement and experimentation are necessary.

Recommendations for LLMs The 5-shot approaches should be preferred for their overall better performance and robustness. Training models with a varied set of examples, including nominal phrases and traps, can improve robustness. Hybrid models should be refined to avoid incorrect references and enhance contextual understanding. The use of more robust rules may lead to better results. Implementing filters to check and correct generated or lost phrases and robust pre-processing techniques can further improve coherence and relevance.

8 Future Directions and Emerging Trends

8.1 Leveraging LLMs for Cataphora Resolution

LLMs, especially those based on transformer architectures like GPT (Generative Pretrained Transformer) [54, 37], Mistral[15], Llama [47] and BERT (Bidirectional Encoder Representations from Transformers) [17], are inherently designed to capture deep contextual cues from text. Their ability to model vast amounts of data allows them to infer relationships and references within text more effectively than earlier models with a more nuanced comprehension and prediction of language patterns.

However, despite their efficacy, LLMs encounter challenges in cataphora resolution, including biases within training data, complexities associated with capturing long-range dependencies in extensive texts, and limitations in fully grasping the subtleties of forward references.

Integrating tasks explicitly focused on forward reference resolution during the training phase could augment the model's capability to address such linguistic structures. This involves incorporating longer contexts with deeper understanding, which can be achieved through techniques such as fine-tuning processes that incorporate high contextual awareness [7]. These approaches may entail fine-tuning existing LLMs on datasets specifically tailored for cataphora resolution or devising pretraining objectives that encompass tasks explicitly related to forward reference resolution.

Furthermore, hybrid models are proposed as an additional strategy, combining the strengths of LLMs with rule-based or symbolic reasoning methods. This hybrid approach is particularly beneficial for scenarios where contextual clues are too subtle for purely statistical models to interpret effectively. The dual-layered processing approach —where the rule-based system first hypothesize potential referents based on context, followed by an LLM evaluating these hypotheses against linguistic rules— ensures both accuracy and adherence to specific linguistic norms. This not only enhances the precision in resolving cataphora but also contributes to the interpretability of the models' decision-making process, a crucial aspect for applications requiring transparent and explainable AI solutions.

However, implementing hybrid models comes with its set of challenges, including the integration of disparate processing paradigms and optimizing the system for efficiency without compromising accuracy.

8.2 Incorporating External Knowledge, Commonsense Reasoning and Chain of Thoughts

The incorporation of external knowledge bases into LLMs through the use of knowledge bases like Wikidata [50], YAGO [45] and ConceptNet [44] offer structured, encyclopedic information that LLMs can leverage to understand the world better. For instance, when a model encounters a forward-reference pronoun, it can query these knowledge bases to fetch relevant information that might indicate possible referents. Integrating these databases requires sophisticated techniques such as entity linking, where entities mentioned in text are matched with their counterparts in the database, and entity disambiguation, to ensure that the model correctly interprets which entity it is dealing with based on the context. This approach not only enriches the model's understanding of specific references but also broadens its general world knowledge, enabling it to make more informed inferences about the text.

This could also involve using entity linking or incorporating world knowledge to provide clues about likely referents in the hybrid models approach proposed [40].

Beyond the integration of factual knowledge, enhancing LLMs with commonsense reasoning capabilities is crucial for resolving cataphora effectively. Commonsense reasoning involves understanding everyday knowledge about the world, human behavior, and the unwritten rules governing interactions and events. By incorporating commonsense databases like COMET-ATOMIC [14] or utilizing models trained on commonsense datasets, LLMs can access a vast array of general knowledge and reasoning patterns. This enhancement allows models to perform beyond mere text pattern recognition, enabling them to infer unstated but implied elements within the text. For example, if a sentence begins with a pronoun followed by an action that typically requires a specific type of entity, commonsense reasoning can help narrow down the referent's identity based on typical roles associated with the action, even before the entity is explicitly mentioned.

The Chain of Thoughts (CoT) [51] can also provide a structured approach to improving model interpretability and reasoning capabilities. CoT was originally designed to improve models' problem-solving abilities by iteratively breaking down complex problems into simpler, sequential steps, can be similarly leveraged for the complex task of cataphora resolution. This involves guiding the model through a series of logical steps or thought processes, enhancing its ability to utilize external knowledge and apply commonsense reasoning in a more transparent and structured manner.

The use of CoT encourages models to explicitly articulate intermediate reasoning steps that lead to the resolution of a cataphoric reference. By doing so, LLMs can be directed to first identify the presence of a forward reference, then consult relevant external knowledge bases or commonsense databases, and finally apply this information to infer the most likely referent. For example, upon encountering a pronoun that refers forward, the model can generate a chain of thoughts like: "The pronoun *she* is used. Based on the context and commonsense reasoning, *she* is likely to refer to a person with a leadership role. Consulting external knowledge suggests *she* could be referring to the CEO mentioned later in the text." This step-by-step approach not only aids in resolving the reference but also makes the model's decision-making process more interpretable.

One significant issue is the seamless integration of external databases without compromising the model's coherence and performance. Additionally, there's the concern of ensuring the relevance and accuracy of the information retrieved from external sources, as well as the computational overhead involved in querying and processing this information in real-time. Moreover, ethical considerations arise regarding the potential biases embedded in external knowledge sources and how they might influence the model's understanding and outputs.

9 Conclusion

9.1 Summary of Key Findings and Insights

The ability of language models to understand and resolve cataphoric references remains a challenging yet fascinating frontier in NLP. The insights from the research paper not only underscore the complexity of cataphora but also chart a course for advancing our models' capabilities in this area.

Embracing Hybrid Models The proposition of hybrid models offers a promising avenue for enhancing cataphora resolution. These models synergize the predictive strengths of statistical language models with the deterministic nature of rule-based systems. By doing so, they can navigate the subtleties of language that purely statistical approaches might overlook, especially in instances where contextual cues are minimal or ambiguous.

Integrating Structured Knowledge for Contextual Enrichment The incorporation of structured external knowledge into language models heralds a

significant leap towards resolving cataphora. By leveraging entity linking and world knowledge, models gain access to a broader context beyond the immediate text, enabling them to infer potential referents with greater accuracy. This approach underscores the importance of context, not just within the text but also in the broader knowledge ecosystem, enriching the model's inference capabilities.

Overcoming Data and Training Hurdles A significant barrier to cataphora resolution is the scarcity of dedicated datasets. The call for diverse and complex linguistic structures in training data is a critical step toward developing models capable of sophisticated resolution strategies. Such data not only enriches the training process but also simulates the multifaceted nature of language, preparing models for the wide array of real-world scenarios they will encounter.

Advancing Pretraining and Fine-Tuning Techniques An approach for improvement lies in refining the pretraining and fine-tuning methodologies of language models. By embedding tasks within the training phase that specifically target the resolution of forward references, models can develop a nuanced understanding of cataphora. This tailored approach to training necessitates a departure from generic pretraining objectives, urging a focused effort on scenarios where models discern and interpret cataphoric references.

Benchmarking and Evaluation: Setting New Standards The development of comprehensive benchmark datasets and nuanced evaluation metrics is imperative for advancing cataphora resolution. These tools will not only facilitate a deeper understanding of model performance but also drive innovation by setting new standards and challenges for researchers and developers in the field.

9.2 Implications for NLP Research and Applications

We hope this exploration of cataphora detection and resolution within the field of NLP will help on shining light on this linguistic phenomena that didn't meet the same interest as anaphoras. Beyond advancing theoretical understanding, this research has profound implications for real-world scenarios, to enhance machines' capacity to comprehend and process natural language effectively. By leveraging contextualized and large language models for cataphora resolution and with the handcraft of rules for hybrid models, NLP researchers are at the forefront of pioneering endeavors. These models not only hold the potential to elevate text comprehension but also to facilitate more nuanced interactions between machines and human language, thereby revolutionizing various domains.

The ramifications of cataphora resolution extend across diverse applications, including information retrieval and search engines, question answering systems, machine translation, and text summarization. In information retrieval, accurate resolution enables systems to provide more relevant search results by deciphering complex linguistic structures. Similarly, in question answering systems, resolving

cataphoric references leads to more precise responses, enhancing user experience. Moreover, in machine translation, contextually appropriate translations are ensured by interpreting pronouns and references accurately. Lastly, in text summarization, coherent and informative summaries are generated through accurate resolution, resulting in clearer and more concise summaries. Despite challenges like handling forward references and data limitations, ongoing advancements in model architectures and training methodologies can be of capital importance for advancing this research field.

9.3 Future Research Opportunities in Cataphora Detection and Resolution

Despite recent advancements, numerous open research questions and challenges persist, paving the way for future investigations in the realm of cataphora detection and resolution. Some potential avenues for further research include:

One prospective direction involves optimizing the integration of Chain of Thoughts (CoT) with external knowledge and commonsense reasoning, with a focus on efficiency and scalability. Streamlining the CoT process could mitigate computational demands while upholding or enhancing the quality of cataphora resolution. Moreover, the development of sophisticated methods for dynamically selecting and integrating external knowledge sources based on the model's ongoing thought processes could bolster performance across a broad spectrum of NLP tasks.

Another avenue ripe for exploration is the development of hybrid models for cataphora resolution, which entail the effective integration of various reasoning and learning methodologies. Future endeavors might concentrate on refining integration techniques to ensure the seamless operation of hybrid systems, investigating their applicability across diverse linguistic and domain-specific contexts, and extending the approach to address other intricate NLP tasks that stand to benefit from a nuanced comprehension of language structures.

References

1. de Almeida, P.M.N.S., Neto, J.F.: The processing of cataphora coreference in brazilian portuguese. ExLing 2020 p. 81 (2020)
2. Barrière, C.: Natural language understanding in a semantic web context. Springer International Publishing, Basel, Switzerland, 1 edn. (nov 201)
3. Boland, J.: The relationship between syntactic and semantic processes in sentence comprehension. Language and Cognitive Processes - LANG COGNITIVE PROCESS 12, 423–484 (08 1997)
4. Chiou, M., Huang, Y.: Np-anaphora in modern greek: A partial neo-gricean pragmatic approach. Journal of Pragmatics 42(7), 2036–2057 (2010), linguistic and Cognitive aspects of Reference
5. Improving Machine Learning Approaches to Coreference Resolution. Association for Computational Linguistics (2002)
6. Cutting, J.: In: Pragmatics and Discourse, pp. 1–80. Routledge (March 2002)

7. Ding, Y., Zhang, L.L., Zhang, C., Xu, Y., Shang, N., Xu, J., Yang, F., Yang, M.: Longrope: Extending llm context window beyond 2 million tokens (2024)
8. Fedele, E., Kaiser, E.: Looking back and looking forward: Anaphora and cataphora in italian (2014), https://api.semanticscholar.org/CorpusID:8301599
9. Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N.F., Peters, M.E., Schmitz, M., Zettlemoyer, L.: Allennlp: A deep semantic natural language processing platform. CoRR abs/1803.07640 (2018), http://arxiv.org/abs/1803.07640
10. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. vol. 3408, pp. 345–359 (04 2005)
11. Hetzer, J.F.: Cross-lingual Coreference Resolution and Neural Machine Translation. Ph.D. thesis, Karlsruhe Institute of Technology (2022)
12. Hoste, V.: Optimization issues in machine learning of coreference resolution. Ph.D. thesis, Universiteit Antwerpen. Faculteit Letteren en Wijsbegeerte. (2005)
13. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. CoRR abs/1902.00751 (2019)
14. Hwang, J.D., Bhagavatula, C., Bras, R.L., Da, J., Sakaguchi, K., Bosselut, A., Choi, Y.: Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs (2021)
15. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023)
16. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., de las Casas, D., Hanna, E.B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L.R., Saulnier, L., Lachaux, M.A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T.L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mixtral of experts (2024)
17. Joshi, M., Levy, O., Zettlemoyer, L., Weld, D.: BERT for coreference resolution: Baselines and analysis. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5803–5808. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
18. Kadam, S., Vaidya, V.: Review and analysis of zero, one and few shot learning approaches. In: Abraham, A., Cherukuri, A.K., Melin, P., Gandhi, N. (eds.) Intelligent Systems Design and Applications. pp. 100–112. Springer International Publishing, Cham (2020)
19. Kazanina, N., Phillips, C.: Differential effects of constraints in the processing of russian cataphora. Quarterly Journal of Experimental Psychology 63(2), 371–400 (2010)
20. Kiziltan, Z., Lippi, M., Torroni, P.: Constraint detection in natural language problem descriptions. In: International Joint Conference on Artificial Intelligence (2016)
21. Kush, D., Dillon, B.: Principle b constrains the processing of cataphora: Evidence for syntactic and discourse predictions. Journal of Memory and Language 120, 104254 (2021)
22. Labrak, Y., Rouvier, M., Dufour, R.: A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks (2023)
23. Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. Computational Linguistics 39(4), 885–916 (12 2013)

24. Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. Computational Linguistics 39(4), 885–916 (12 2013)
25. Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D.: Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. pp. 28–34. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011)
26. Lee, H., Surdeanu, M., Jurafsky, D.: A scaffolding approach to coreference resolution integrating statistical and rule-based models. Natural Language Engineering 23(5), 733–762 (2017)
27. Lee, K., He, L., Lewis, M., Zettlemoyer, L.: End-to-end neural coreference resolution. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 188–197. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017)
28. Lenat, D.B., Guha, R.V., Pittman, K., Pratt, D., Shepherd, M.: Cyc: toward programs with common sense. Commun. ACM 33(8), 30–49 (aug 1990)
29. Addressing Lexical and Semantic Ambiguity in Natural Language Requirements (2018)
30. Liddy, E.D.: Anaphora in natural language processing and information retrieval. Information Processing Management 26(1), 39–52 (1990), special Issue: Natural Language Processing and Information Retrieval
31. Liu, Z., Lin, Y., Sun, M.: World Knowledge Representation, pp. 163–216. Springer Nature Singapore, Singapore (2020)
32. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60 (2014)
33. Min, B.: Exploring pre-trained transformers and bilingual transfer learning for Arabic coreference resolution. In: Ogrodniczuk, M., Pradhan, S., Poesio, M., Grishina, Y., Ng, V. (eds.) Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference. pp. 94–99. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021)
34. Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J.: Large language models: A survey. arXiv preprint arXiv:2402.06196 (2024)
35. Mitkov, R.: Anaphora Resolution. Studies in Language and Linguistics, Routledge, London, England (2002)
36. Mitkov, R., Evans, R., Orăsan, C., Ha, L.A., Pekar, V.: Anaphora resolution: To what extent does it help nlp applications? In: Anaphora: Analysis, Algorithms and Applications, pp. 179–190. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
37. OpenAI: Gpt-4 technical report (2024)
38. Otmazgin, S., Cattan, A., Goldberg, Y.: F-coref: Fast, accurate and easy to use coreference resolution (2022)
39. Poot, C., van Cranenburgh, A.: A benchmark of rule-based and neural coreference resolution in dutch novels and news. CoRR (Nov 2020)
40. Qiao, S., Fang, R., Zhang, N., Zhu, Y., Chen, X., Deng, S., Jiang, Y., Xie, P., Huang, F., Chen, H.: Agent planning with world knowledge model (2024)
41. Recasens, M., Màrquez, L., Sapena, E., Martí, M.A., Taulé, M., Hoste, V., Poesio, M., Versley, Y.: SemEval-2010 task 1: Coreference resolution in multiple languages.

In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 1–8. Association for Computational Linguistics, Uppsala, Sweden (Jul 2010), https://aclanthology.org/S10-1001
42. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A machine learning approach to coreference resolution of noun phrases. Computational Linguistics 27(4), 521–544 (12 2001)
43. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A machine learning approach to coreference resolution of noun phrases. Comput. Linguist. 27(4), 521–544 (dec 2001)
44. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge. CoRR abs/1612.03975 (2016)
45. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A Core of Semantic Knowledge. In: 16th International Conference on the World Wide Web. pp. 697–706 (2007)
46. Sukthanker, R., Poria, S., Cambria, E., Thirunavukarasu, R.: Anaphora and coreference resolution: A review. Information Fusion 59, 139–162 (2020)
47. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023)
48. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models (2023)
49. Trnavac, R., Taboada, M.: Cataphora, backgrounding and accessibility in discourse. Journal of Pragmatics 93, 68–84 (2016)
50. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM 57(10), 78–85 (sep 2014)
51. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023)
52. Wong, K., Maruf, S., Haffari, G.: Contextual neural machine translation improves translation of cataphoric pronouns (2020)
53. Xia, P., Durme, B.V.: Moving on from ontonotes: Coreference resolution model transfer (2021)
54. Yenduri, G., M, R., G, C.S., Y, S., Srivastava, G., Maddikunta, P.K.R., G, D.R., Jhaveri, R.H., B, P., Wang, W., Vasilakos, A.V., Gadekallu, T.R.: Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions (2023)
55. Yu, W., Wu, L., Deng, Y., Mahindru, R., Zeng, Q., Guven, S., Jiang, M.: A technical question answering system with transfer learning. pp. 92–99 (01 2020)