Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 15-27 recibido 13-02-2024 revisado 09-05-2024 aceptado 15-05-2024

Are rule-based approaches a thing of the past? The case of anaphora resolution

¿Son los métodos basados en reglas cosa del pasado? El caso de la resolución de la anáfora

Ruslan Mitkov,[1] **Le An Ha**[2]
[1]Lancaster University
[2]Ho Chi Minh City University of Foreign Languages and Information Technology
[1]r.mitkov@lancaster.ac.uk, [2]anhl@huflit.edu.vn

Abstract: In this paper, we evaluate and compare new variants of a popular rule-based anaphora resolution algorithm with the original version. We seek to establish whether configurations t hat b enefit fr om De ep Le arning, LL Ms an d eye-tracking data (always) outperform the original rule-based algorithm. The results of this study suggest that while algorithms based in Deep Learning and LLMs usually perform better than rule-based ones, this is not always the case, and we argue that rule-based approaches still have a place in today's research.
Keywords: Anaphora resolution, Rule-based, deep learning.

Resumen: En este artículo evaluamos y comparamos nuevas variantes de un conocido algoritmo de resolución de anáforas basado en reglas con la versión original. Buscamos establecer si los enfoques que se benefician de aprendizaje profundo, grandes modelos de lenguaje (LLMs) y datos de eye-tracking (siempre) superan al algoritmo original basado en reglas. Los resultados de este estudio sugieren que, aunque los algoritmos basados en aprendizaje profundo y grandes modelos de lenguaje suelen rendir mejor que los basados en reglas, no siempre es así. Por lo tanto, sostenemos que los enfoques basados en reglas siguen teniendo cabida en la investigación actual.
Palabras clave: Resolución de Anáforas, Basado en Reglas, Aprendizaje profundo.

1 Rationale

Back in the 1990s there was a dramatic surge in the development of operational anaphora resolution algorithms due among other things, to the availability of part-of-speech taggers and parsers as well as corpora which inspired research into the development and evaluation of knowledge-poor approaches. Initially the algorithms were predominantly rule-based where specific rules were proposed as a result of observing specific patterns in the corpus data (Lappin and Leass, 1994; Mitkov, 1998; Baldwin, 1998). While influential, the rule-based approaches gradually gave a way to statistical approaches (Ge, Hale, and Charniak, 1998) and later to machine learning approaches which dominated the research on anaphora resolution in the early 2000s. More recently data-driven models and, most prominently, Deep Learning and Large Language Models, have taken the world by storm. Deep Learning is used almost everywhere, in almost every discipline and Natural Language Processing (NLP) is not an exception. The widespread use of Deep Learning approaches and Large Language Models (LLMs) in NLP has not bypassed the task of anaphora resolution and several recent studies have been reported in the last five years or so (see section 2.3).

Deep Learning (DL) has been highly successful so far indeed, with improvements reported for almost every NLP task and application. However, as seen on numerous occasions, the outputs of DL models are not always perfect, with the failure of Neural Machine Translation to successfully translate multiword expressions being an obvious

ISSN 1135-5948 DOI 10.26342/2024-73-1

example (Colson, 2019; Mitkov, 2019)[1] . In addition, there have been earlier studies which report that machine learning approaches to anaphora resolution do not fare necessarily better than the 'old-fashioned' rule-based ones (Stuckardt, 2002; Stuckardt, 2003; Stuckardt, 2005).

More recently, Large Language Models (LLMs) has been another influential development with many NLP studies following the lead of, and comparing their performance with, generative models.

In this study we seek to establish (whether and) to what extent up-to-date Machine Learning, Deep Learning and LLM approaches as well as approaches benefiting from modern resources such as eye-tracking gaze data, deliver better results than rule-based approaches developed in the 1990s when the above techniques and resources were not available. To this end we take the extensively used Mitkov (1998)'s knowledge-poor approach as a testbed and build several modifications of the original algorithm which incorporate popular recent (statistical, machine learning, deep learning and LLM) techniques and benefit from gaze data. Then we compare the performance of the enhanced systems and seek to establish whether and to what extent they outperform the original algorithm, and which version fares best.

The rest of the paper is structured as follows. Section 2 provides the preliminaries of this research by introducing Mitkov's knowledge-poor approach to anaphora resolution used here as a testbed for our experiments, discusses the use of eye-tracking in anaphora resolution, and outlines recent work on Deep Learning and LLMs for anaphora and coreference resolution. Section 3 covers the data used and methodology employed by providing information on the annotated data prepared for this study and the methodology adopted. Section 4 elaborates on the experiments conducted and comments on the results obtained. Finally, the last section offers the concluding remarks for this study. Code and data used in the experiments will be made available on github.

[1]The performance of NMT systems has significantly improved since these studies were conducted and published.

2 Preliminaries

2.1 Mitkov's knowledge poor pronoun resolution approach as testbed

(Mitkov, 1996; Mitkov, 1998)'s robust, knowledge-poor approach to pronoun resolution was motivated by the pressing need in the 1990s for anaphora resolution algorithms operating robustly in real-world, knowledge-poorer environments in order to meet the demands of practical NLP systems. The first version of the algorithm was reported in (Mitkov, 1996) as an inexpensive, fast and yet reliable alternative to the labour-intensive and time-consuming construction of a knowledge-based system[2]. This project was also an example of how anaphors can be resolved quite successfully (at least in a specific genre, namely computer/technical manuals) without any sophisticated linguistic knowledge or even without parsing. In addition, the evaluation showed that the basic set of factors (referred to as 'indicators', see below) employed can work well not only for English, but also for other languages including French, Bulgarian, Polish and Arabic (Mitkov and Stys, 1997; Mitkov and Belguith, 1998; Mitkov, Belguith, and Stys, 1998; Mitkov and Barbu, 2000; Tanev and Mitkov, 2002; Mitkov, 2006).[3]

Mitkov's approach relies on a list of preferences known as antecedent indicators. The approach operates as follows: it works from the output of a text processed by a part-of-speech tagger and an NP extractor, identifies noun phrases which precede the anaphor within a distance of 2 sentences[4], checks them for gender and number agreement with the anaphor and then applies the indicators to the remaining candidates by assigning a positive or negative score (2, 1, 0 or -1). The noun phrase[5] with the highest composite sco-

[2]The approach has become better known through a later updated publication (Mitkov, 1998).

[3]In fact the performance reported for Slavonic languages was higher than for English due to the fact that they have a three-gender system: gender agreement would filter many ineligible candidates.

[4]Subsequent versions of the approach have used search scopes of different lengths (2, 3 or 4 sentences), but the original algorithm only considered two sentences prior to the sentence containing the anaphor. The NP patterns are limited to the identification of base NPs and do not include complex or embedded phrases.

[5]The approach handles only pronominal anaphors

re is proposed as antecedent.

The antecedent indicators are applied to all NPs which have passed the gender and number filters.[6] These indicators can act in either a boosting or an impeding capacity. The boosting indicators apply a positive score to an NP, reflecting a positive likelihood that it is the antecedent of the current pronoun.

In contrast, the impeding ones apply a negative score to an NP, reflecting a lack of confidence that it is the antecedent of the current pronoun. Most of the indicators are genre-independent and related to coherence phenomena (such as salience and distance) or to structural matches, whereas others are genre-specific.[7] The boosting and impeding indicators are described in detail in Mitkov (1998). The work presented in Mitkov, Evans, and Orasan (2002) provides some additional detail on the indicators used by the algorithm.

The aforementioned antecedent indicators are preferences and not absolute factors. There might be cases where one or more of the antecedent indicators do not 'point' to the correct antecedent. For instance, in the sentence 'Insert the cassette into the VCR making sure it is turned on', the indicator prepositional noun phrases would penalise the correct antecedent. When all preferences (antecedent indicators) are taken into account however, the right antecedent is still likely to be tracked down - in the above example, the prepositional noun phrases heuristic stands a good chance of being overturned by the collocation match heuristics since the collocation 'The VCR is turned on' is likely to appear previously in the text, being typical of video technical manuals.

The antecedent indicators have proved to be reasonably efficient in identifying the right antecedent and the results show that for the genre of technical manuals they may be no less accurate than syntax- and centering-based methods (see Mitkov 1998). The approach is not dependent on any theories or assumptions; in particular, it does not operate on the assumption that the subject of the previous utterance is the highest-ranking candidate for the backward-looking center - an approach which can sometimes lead to incorrect results.[8]

Mitkov's original algorithm was enhanced and developed into the fully-automatic pronoun resolution system referred to as MARS (Mitkov, Evans, and Orasan, 2002)[9]. The initial implementation of MARS employed the FDG shallow parser as its main preprocessing tool and was based on a revised version of the original algorithm.

The initial implementation of MARS followed Mitkov's original approach closely, the main differences being (i) the addition of three new indicators and (ii) the change in the way some of the indicators were implemented or computed due to the available preprocessing tools. In its later version, MARS also used a program for automatically recognising instances of anaphoric or pleonastic pronouns (Evans, 2001) and intrasentential syntax filters.

The system operated in five phases. In phase 1, the text to be processed is parsed syntactically, using Conexor's FDG Parser (Tapanainen and Jarvinen, 1997) which returns the parts of speech, morphological lemmas, syntactic functions, grammatical number, and dependency relations between tokens in the text which facilitates complex noun phrase (NP) extraction.

In phase 2, anaphoric pronouns are identified and non-anaphoric and non-nominal instances of *it* are filtered using the machine learning method described in Evans (2001).

In phase 3, for each pronoun identified as anaphoric, candidate NPs are extracted from the heading of the section in which the pronoun appears, and from the current and preceding two sentences within the paragraph under consideration. Once identified, these candidates are subjected to further morphological and syntactic tests. Extracted candi-

whose antecedents are noun phrases.

[6]The approach takes into consideration the fact that in English there are certain collective nouns which do not agree in number with their antecedents (e.g. government, team, parliament etc. can be referred to by 'they'; equally some plural nouns such as data can be referred to by 'it') and are thus exempted from the agreement test.

[7]Typical of the genre of user guides.

[8]For instance, subject-favouring methods or methods heavily relying on syntactic parallelism would incorrectly propose the utility as the antecedent of it in the sentence 'The utility shows you the LIST file on your terminal for a format similar to that in which it will be printed' as it would prefer the subject as the most salient candidate. The indicating verbs preference of Mitkov's approach, however, would prefer the correct antecedent the LIST file.

[9]MARS stands for Mitkov's Algorithm to pronoun ReSolution.

dates are expected to obey a number of constraints if they are to enter the set of competing candidates, i.e. the candidates that are to be considered further. Competing candidates are required to agree with the pronoun in terms of number and gender, as is the case in the original algorithm. They must also obey syntactic constraints (Mitkov, Evans, and Orasan, 2002). In phase 4, 14 preferential and impeding factors are applied to the sets of competing candidates. On application, each factor applies a numerical score to each candidate, reflecting the extent of the system's confidence about whether the candidate is the antecedent of the current pronoun. In the implemented system, certain practical issues led to the weights assigned by indicators being computed in a different way from that described in the original algorithm. The full details of these differences are beyond the scope of the current paper, but they are described in detail in (Mitkov, Evans, and Orasan, 2002). In addition, three new indicators were added, one of which (syntactic parallelism) exploits new, previously unavailable features of the pre-processing software.

Finally, in phase 5, the candidate with the highest composite score is selected as the antecedent of the pronoun. Ties are resolved by selecting the most recent highest scoring candidate.

Further versions of MARS incorporated several advancements over the system described in Mitkov, Evans, and Orasan (2002). These improvements covered the inclusion of more precise and strict number and gender agreement, and the addition of one indicator employing the modelling of selectional restrictions.

Finally, MARS was improved to cater for several frequent causes of apparent number disagreement. These consist of (i) collective nouns, (ii) gender under-specification, (iii) quantified nouns/indefinite pronouns, and (iv) organisation names by NER. These cases were handled by a combination of gazetteers and the integration of animacy recognition modules (Orasan and Evans, 2001) and named entity recognition (Cunningham et al., 2000). Patterns were used to identify the occurrence of quantified NPs in the parsed text. MARS's recognition of the gender of NP candidates was improved. In addition to gazetteers, an NER system was used to recognise person names and a system for animacy recognition deployed.

All versions of Mitkov's knowledge-poor algorithm which have been widely used in different anaphora resolution studies were rule-based. Versions enhanced by Machine / Deep Learning techniques have been developed (and are reported) for the first time (in this paper) in line with the objectives of this study.

2.2 Pronoun resolution and eye-movement measures

Eye tracking is a process where an eye-tracking device measures the point of gaze of an eye (gaze fixation) or the motion of an eye (saccade) relative to the head and a computer screen (Duchowski, 2017). Different eye-tracking measures (usually divided into early and late) are indicative of different aspects of cognitive processing. For example, early gaze measures such as first fixation duration give information about the early stages of lexical access and syntactic processing, while late gaze measures such as total dwell time or total number of fixations give information about processes such as textual integration, syntactic and semantic processing and disambiguation. A series of studies on eye tracking during reading show that gaze data is sensitive to linguistic phenomena such as word frequency, verb complexity and lexical ambiguity, as well as contextual effects on word perception (Rayner et al., 2012; Rayner, 1998; Rayner and Duffy, 1986).

Eye-movement measures have been used in psycholinguistic studies of pronoun resolution. The most relevant study to the present research is conducted by Foraker and McElree (2007), who found that resolution occurred earlier for he/she pronouns than for the it pronoun in both clefted[10] and non-clefted sentences. In a follow-up experiment they used eye tracking to identify whether the time-course difference could be due to the ambiguity of the it pronouns or to the antecedent being actively maintained in the focal span of the readers. They hypothesised that if the reason for the time-course difference was in the active state of the antecedents, the differences would be observed in measures such as first-pass duration. Alter-

[10] Constructions of this type include 'It is Jamie for whom we are looking' or 'And most disturbing, it is educators, not students, who are blamed for much of the wrongdoing' (Li et al., 2009).

natively, if the reason was the ambiguity of the it pronoun, the differences would be observed in measures signalling reanalysis, such as increased immediate regressions from the coreference region, longer second-pass times and regression path durations. Their analysis indicated that conditions with the pronoun *it* caused more first-pass regressions than those with *he/she*, which lead the authors to conclude that the it-pronoun conditions were in fact functionally ambiguous. This ambiguity was later resolved by looking at the token occurring after the pronoun, which provided "immediate diagnostic information that helped identify and repair the coreference bond" (Foraker and McElree, 2007). No effects of clefting were found. The experiments of this study provide convincing evidence that the locus of the time-course difference was the it-pronoun owing to its ambiguity.

Based on these results, we hypothesise that gaze data contains traces of the way humans perform pronoun resolution and that these traces can be used to improve the performance of automatic coreference resolution. This approach has been previously used in other areas of NLP, including classifying referential and non-referential it (Yaneva et al., 2018), sentiment analysis (Rotsztejn, 2018), part-of-speech tagging (Barrett et al., 2016), and multiword expressions (Rohanian et al., 2017), among others.

2.3 Recent work on Deep Learning and LLMs for anaphora and coreference resolution

The employment of Deep Learning (DL) for a number of NLP tasks and applications[11] has been an important trend in recent years and there has been hardly any NLP area in which Deep Learning methods have not been made use of. Anaphora resolution as a crucial NLP task has not gone unnoticed by researchers who have been experimenting with and applying DL approaches in the hope of improving performance.

In one of the first studies employing deep learning for anaphora/coreference, Clark and Manning (2015) described a coreference resolution system based on neural networks which automatically learned dense vector representations for mention pairs. These were derived from distributed representations of the words in the mentions and surrounding context and captured semantic similarity which could assist the coreference resolution process. The representations were used to train an incremental coreference system which can exploit entity-level information.

Clark and Manning (2016) applied reinforcement learning to optimise a neural mention-ranking model for coreference evaluation metrics. The authors experimented with two approaches: REINFORCE policy gradient algorithm and a reward-rescaled max-margin objective. They found the latter to be more effective, resulting in a significant improvement over the state of the art on the English and Chinese portions of the CoNLL 2012 Shared Task.

Wiseman, Rush, and Shieber (2016) employed recurrent neural networks (RNNs) to learn latent global representations of entity clusters directly from their mentions. They showed that such representations are especially useful for the prediction of pronominal mentions and can be incorporated into an end-to-end coreference system which outperformed the state of the art without requiring any additional search.

More recently, Plu et al. (2018) presented an improved version of the Stanford 'deep-coref' system by enhancing it with semantic features, and reported a minimal increase of the F-score, while Sukthanker et al. (2018) described an entity-centric neural crosslingual coreference model which builds on multi-lingual embeddings and language-independent features and performs well in intrinsic and extrinsic evaluations.

Other recent work which employs deep learning for anaphora and/or coreference resolution include Meng and Rumshisky (2018) who used a triad-based neural network system to generate affinity scores between entity mentions for coreference resolution, and Nitoń, Morawiecki, and Ogrodniczuk (2018) who experimented with several configurations of deep neural networks for coreference resolution in Polish.

Latest research on using deep learning models for coreference resolution is summarised in Liu et al. (2023).

[11] See Mitkov (2003; Mitkov (2022) for distinctions between NLP tasks and NLP applications where the former include part-of-speech tagging, parsing, word sense disambiguation, semantic role labelling, anaphora resolution, etc., and the latter include machine translation, text summarisation, text categorisation, information extraction and question answering, among others.

Latest Large Language Models have been explored in anaphora resolution as well. Yang et al. (2022) study how well ChatGPT-like models do on anaphora resolution. They conclude that LLMs perform poorly on the task of coreference resolution without fine-tuning. These models achieve relatively better performance on pronouns and mention pairs with high similarity. The authors also reported that the capabilities of such models to identify coreferent mentions are limited and prompt-sensitive, leading to inconsistent results.

In another recent study, Vadász (2023) reports experiments on using ChatGPT for pronoun resolution in Hungarian. The experiments suggest that while ChatGPT does reasonably well, it is far away from the ideal performance.

3 Data and methodology

3.1 Annotated corpus

In this study we exploited two existing corpora of English text annotated with information about the eye movements of readers, recorded using eye tracking equipment. These were:

- The 51,254-token English portion of the Dundee corpus (Kennedy, Hill, and Pynte, 2003), a collection of news articles from the Independent. The eye tracking data encoded in this corpus was recorded from ten English-speaking readers using a Dr Bouis Oculometer Eyetracker with a 1 kHz monocular (right) sampling rate.

- The 56,419-token Ghent Eye-Tracking Corpus (GECO) (Cop et al., 2016), which comprises the text of the Agatha Christie novel The Mysterious Affair at Styles. It is annotated with eye tracking data recorded from English monolinguals and Dutch-English bilinguals using a tower-mounted EyeLink 1000 system with a sampling rate of 1 kHz.

To the eye-tracking information encoded in the GECO and Dundee corpora, we added a second annotation layer to encode information about examples of the pronoun it occurring in these texts. This includes information about the locations of the pronouns and the candidate NPs preceding them in the same sentence and in the two preceding sentences. Due to the laboriousness of the task for human annotators of manually identifying the sets of NPs preceding each pronoun in a text, we implemented a semi-automatic annotation task. We used the Charniak parser (Charniak, 2000) to automatically identify and list an initial set of preceding NP candidates for each example of the pronoun *it*. These lists were then post-edited by the human annotators who deleted incorrect candidates and inserted any NPs omitted due to parsing errors. When post-editing the lists, the human annotators also enforced number and gender agreement constraints to ensure that the lists only contained antecedent candidates that are singular in number and neuter in gender. The identification of these attributes is one in which automatic methods are currently unreliable. Table 1 presents information on these characteristics of the corpora.

3.2 Methodology

We seek to establish to what extent (and whether) up-to-date Machine Learning, Deep Learning and LLM techniques as well as the exploitation of eye-tracking gaze data, could deliver better results than old-fashioned and popular-in-the-1990s rule-based approaches. In this particular study Mitkov's knowledge-poor approach was used as a testbed. Several modifications of the original algorithm (Mitkov, 1998) incorporating popular recent statistical, machine learning and deep learning techniques as well as resources were implemented and the performance of the enhanced algorithms was compared with that of the original algorithm. In this study the values of the antecedent indicators[12] were regarded as features whose weights were to be optimised. In addition to features derived from original antecedent indicators, each NP candidate was associated with a set of language model features and a set of gaze features.[13]

For each candidate of a particular pronoun, the sentence containing the pronoun was identified and a variant sentence was generated in which the pronoun was replaced by the candidate NP with the probability of each variant sentence encoded through language model features. This particular experiment employed 14 language models derived from deep learning and vector representations of words/characters, presented by Trinh and Le (2018). These 14 models differ from each other with respect to their neural net-

[12] See Mitkov (1998, 2002) for description of the antecedent indicators.

[13] A variety of gaze features encoded for each token in the Dundee and GECO corpora, was exploited.

work settings and training data used with three of them being similar to those used by ELMo (Peters et al., 2018) a word embedding method which fared better than word2vec.

In addition, information about a variety of gaze features encoded for each token in the Dundee and GECO corpora, was exploited. The gaze features used were First Fixation Duration (The duration of the first fixation to fall on the token), Second Fixation Duration: (The duration of the second fixation to fall on the token), Last Fixation Duration (The duration of the final fixation to fall on the token), Have second fixation (Readers fixate on the token twice or more), Fixation Count (The number of times readers fixate on the token) and Skip Rate (The proportion of times that the token is not fixated upon by readers).

Source	**#W**	**#Pr**	**#Can**	**PP**
GECO	56,419	533	2,391	4.5
Dundee	51,254	224	3,248	14.5
TOTAL	107,673	757	5,639	7.4

Tabla 1: Characteristics of the annotated corpus (#W: number of words, #Pr: Number of pronouns, #Can: Number of candidate antecedents, PP: Number of singular neuter candidates per pronoun)

4 Experiments and results

After deriving the values of the full feature set (comprising the original MARS antecedent indicator features, the DL language model features, and the gaze features), we conducted experiments to optimise the weights of various combinations of features in 10-fold-cross-validation and cross-corpus evaluation settings. In 10-fold cross-validation, we evaluated over the corpora by exhaustively training on nine tenths (folds) of them and testing on the remaining tenth (fold). In cross-corpus evaluation, we evaluated by training on one whole corpus (e.g. GECO) and testing on the other (e.g. Dundee).

We experimented with various machine learning models trained to predict whether a candidate NP is the actual antecedent of a pronoun (1) or not (0), using our annotated training data. The trained models were then applied to generate a score for each of the candidate NPs in the testing data. Simple linear regression using the least squares approach led to the derivation of most accurate anaphora resolution models, regardless of the evaluation setting (cross-validation or cross-corpus). The evaluation results show that optimisation of antecedent indicator weights can improve accuracy of anaphora resolution model by around 10 % which is statistically significant. In the statistical linear regression model, Deep Learning language model features were included as variables; annotated data used to learn optimal weights for variables.

	Corpus		
Settings	Dundee	GECO	ave.
Antecedent indicators only			
Original	0.39	0.51	0.47
Optimised using			
10-fold cross-val	0.48	0.58	0.55
Cross-corpus	0.43	0.58	0.53

Tabla 2: Effects of optimisation of indicator weights

In the statistical linear regression model, we included the DL language model features as variables and used our annotated data to learn the optimal weights for those variables. Optimising the weighting of the DL language model features improved the accuracy of the anaphora resolution process.

Table 3 shows the effects of the different configurations on the accuracy of anaphora resolution. We compare the accuracy of the original algorithm with various enhanced algorithms corresponding to these configurations. In the table, corpus, absolute refers to the actual number of pronouns correctly resolved by each configuration while corpus, accuracy is the ratio of corpus, absolute to the total number of referential occurrences of it in the corpus. The configurations are based on various combinations of features including the antecedent indicators, DL language model features and gaze features. In our experiments, features are combined by concatenating them into a single vector. The language models used to obtain the language model features were built using deep learning methods applied to huge corpora by Trinh and Le (2018). The results presented here are derived from experiments conducted in the cross-corpus evaluation setting. We consider that the cross-corpus setting better reflects real world scenarios where models should not

	Corpus, absolute count			Corpus, accuracy		
Setting	GECO	Dundee	All	GECO	Dundee	Wt. ave
DeepLM token labelling (deberta-v3-large, cross corpus train/test, best among candidates)	349	149	490	0.686	0.668	0.680**
DeepLM token labelling (deberta-v3-base, cross corpus train/test, best among candidates)	337	143	480	0.662	0.664	0.656**
ChatGPT-4, choose among candidates	309	150	459	0.607	0.672	0.627**
Antecedent Indicators (Original Weights) + DL Language Model Features	301	118	419	0.591	0.529	0.572**
Antecedent Indicators (Original Weights) + DL Language Model Ensemble[1]	302	115	417	0.593	0.516	0.570**
Antecedent Indicators (Original Weights) + Gaze Selected (Dundee)	299	115	414	0.587	0.516	0.566**
ChatGPT-4, no candidate	290	118	408	0.570	0.529	0.560*
Antecedent Indicators (Original Weights) + Gaze All + DL Language Models Ensemble	301	106	407	0.591	0.475	0.556*
NeuralCoref	288	118	406	0.566	0.529	0.554*
Antecedent Indicators + DL Language Models Ensemble	304	98	402	0.597	0.439	0.549*
Antecedent Indicators (Original Weights) + Gaze All	297	105	402	0.583	0.471	0.549*
Antecedent Indicators + Gaze All	297	104	401	0.583	0.466	0.548*
Antecedent Indicators + DL Language Model Features (Original Weights)	299	102	401	0.587	0.457	0.548*
Antecedent Indicators (Optimized Weights)	298	97	395	0.585	0.435	0.540
DeepLM token labelling (deberta-v3-large, cross corpus train/test, raw span)	255	136	391	0.501	0.610	0.534
DeepLM token labelling (deberta-v3-base, cross corpus train/test, raw span)	240	122	362	0.472	0.547	0.494
Antecedent Indicators (Original Weights)	270	90	360	0.530	0.404	0.492
DL Language Models Ensemble[14]	274	70	344	0.538	0.314	0.470
DL Language Model Features	212	60	272	0.417	0.269	0.372
Gaze All	221	49	270	0.434	0.220	0.369
Gaze Selected (GECO)	216	50	266	0.424	0.224	0.363
Gaze Selected (Dundee)	195	42	237	0.383	0.188	0.324

Tabla 3: Effects of different combinations on accuracy. *: statistically significant, McNeymar test, at p¡0.05 when compared with the Antecedent Indicators (Original Weights) configuration. **: statistically significant at p¡0.01 when compared with the Antecedent Indicators (Original Weights) configuration.

be domain-specific.

Table 3 refers to several feature sets and settings:

- Gaze All is the set First Fixation Duration, Second fixation duration, Last Fixation Duration, Have second fixation, Fixation Count, And Skip Rate.
- Gaze Selected (Dundee) is the set Have Second Fixation, Skip rate.
- Gaze selected (GECO) is the set First Fixation Duration, Second fixation duration
- DL Language Model Features is the set of 14 sentence probabilities obtained when the 14 language models presented by Trinh and Le (2018) are used to ob-

tain the probabilities of variant sentences. In this context, variant sentences are versions of the sentence containing the pronoun occurs in which the pronoun has been replaced by the candidate NP.

- DL Language Models ensemble comprises the joint probability of the aforementioned language model features.
- Antecedent Indicators (Original Weights): is the sum of the scores assigned by the antecedent indicators presented in Section 2.1, whose weights were set empirically, in accordance with the original statement of Mitkov's anaphora resolution algorithm (Mitkov, 2002).
- Antecedent Indicators (Optimised Weights): is the sum of the scores assigned by the antecedent indicators briefly outlined in Section 2.1, whose weights were set using the linear regression optimisation method.
- NeuroCoref: In this approach, we employ the HuggingFace NeuroCoref model as it comes pre-packaged. Specifically, we examine NeuroCoref's suggestions for antecedents and compare them against the annotated gold antecedents. This allows us to determine if the suggested antecedent aligns with the expected reference according to the provided annotations.
- DeepLM token labelling: In this approach, we consider the task of anaphora resolution as token labelling. We use the annotated data to finetune a generic language model (in our case, either deberta-v3-base or deberta-v3-large) to the task of identifying the spans that correspond to the annotated antecedents. The proposed antecedent then either the span whose tokens' probabilities that they belong to an antecedent are greater than 0.5 (raw span), or the span whose average tokens' probabilities that they belong to an antecedent is the largest among the candidates provided by the preprocessing steps (best among candidates).
- ChatGPT-4 no candidate. We use ChatGPT to determine the antecedents of *it* using zero-shot learning method. Specifically, we use the following template: "in the following paragraph: {text}, 'it' in '{feature_text}' refers to, give me the exact phrase". The {text} is the whole context in which the pronoun *it* can be found, and the {feature_text} is the part of the text starting with the *it* pronoun. We then get the results from ChatGPT-4 through openAI API, and match the results with the known correct antecedents.
- ChatGPT-4 Choose among candidates. We use ChatGPT to choose the antecedents of *it* among candidates using zero-shot learning method. Specifically, we use the following template: "in the following paragraph: {text}, 'it' in '{feature_text}' refers to (A) {candidate_1} B {candidate_2} ...". The {text} is the whole context in which the pronoun *it* can be found, and the {feature_text} is the part of the text starting with the 'it' pronoun. The list of candidates is provided by the preprocessing steps. We then get the results from ChatGPT-4 through openAI API, and match the results with the known correct antecedents. [15]

5 Discussion

Closer examination of the results (Table 3) allows us to make the following observations.

1. Optimisation of antecedent indicators

   The weights on antecedent indicators, initially set empirically, can be optimised further, but when combined with DL language models and gaze features, the empirically set weights seem to work very well across domains.

2. Deep Learning models

   The inclusion of features derived using language models (deep learning and advanced vector representations) improves the accuracy of the anaphora resolution method.

   In recent years, deep learning language models are getting very good at tasks such as token labelling, and if they are

[15] Unlike the experiments with Deep Learning models which are built on top of the original anaphora resolution algorithm, in both ChatGPT experiments ('ChatGPT-4, no candidate' and 'ChatGPT-4, choose among candidates'), ChatGPT's operation is based on its original methodology due to the impossibility to change the way it operates.

fine-tuned on a small amount of data, they can produce very good results.

3. Large Language Models

   ChatGPT, when asked to identify antecedents on its own performs competitively as compared to (or event better) than selected Deep Learning methods but cannot beat the original algorithm when combined with gaze data or DL Language Model Features or Language Model Ensemble and still is far behind the best performing DeepLM token labelling methods, when asked to predict the exact antecedents. It performs well when asked to pick the correct antecedents among the candidates.

4. Gaze features

   The inclusion of gaze features improves the accuracy of the anaphora resolution method.

   Gaze features do not provide additional information to models that use DL language model features. An interesting observation from our experiments is that the two sources of information appear to be redundant rather than complementary.

The results of the experiments suggest that the rule-based anaphora resolution does not always perform less successfully than models based on deep learning and gaze data – in fact, in some cases it delivers better results. While data-driven approaches generally fare better in these experiments, they still have some way to go in order to be comfortably beat old-fashioned rule-based approaches. The results also show that the weights on antecedent indicators, initially set empirically, can be optimised further. When combined with DL language models and gaze features, empirically set weights appear to work very well across domains. In particular the inclusion of gaze features also improves accuracy of the anaphora resolution and the inclusion of features derived using language models (deep learning and advanced vector representations) also improves accuracy of anaphora resolution. However, gaze features do not provide additional information to models that use DL language model features.

6 Conclusions

Anaphora resolution is arguably one of the most difficult NLP tasks. The results of this study provide evidence that while in most cases Deep Learning and Large Language Models applied to this task show superior performance to rule-based methods, the old-fashioned, rule-based algorithms still perform competitively and still have a place in today's research. An interesting follow-up study might be one that seeks to establish how rule-based algorithms could be better integrated in or combined with new Large Language Models and what the limits of anaphora resolution are. Improvements of the performance of anaphora resolution have been only incremental over the years. Now, with the power of the latest LLMs, how far can we go?

References

Baldwin, B. 1998. Coreference as the foundations for link analysis over free text databases. In *Content Visualization and Intermedia Representations (CVIR'98)*.

Barrett, M., J. Bingel, F. Keller, and A. Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–584, Berlin, Germany, August. Association for Computational Linguistics.

Charniak, E. 2000. A maximum-entropy-inspired parser. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Clark, K. and C. D. Manning. 2015. Entity-centric coreference resolution with model stacking. In C. Zong and M. Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China, July. Association for Computational Linguistics.

Clark, K. and C. D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In J. Su, K. Duh, and X. Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in*

Natural Language Processing, pages 2256–2262, Austin, Texas, November. Association for Computational Linguistics.

Colson, J.-P. 2019. Multi-word units in machine translation: why the tip of the iceberg remains problematic – and a tentative corpus-driven solution. In *MUMTT2019, The 4th Workshop on Multi-word Units in Machine Translation and Translation Technology.*

Cop, U., N. Dirix, D. Drieghe, and W. Duyck. 2016. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49, 05.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan, and Y. Wilks. 2000. Experience using GATE for NLP R&D. In R. Zajac, editor, *Proceedings of the COLING-2000 Workshop on Using Toolsets and Architectures To Build NLP Systems*, pages 1–8, Centre Universitaire, Luxembourg, August. International Committee on Computational Linguistics.

Duchowski, A. T. 2017. *Eye Tracking Methodology: Theory and Practice.* Springer Publishing Company, Incorporated, 3rd edition.

Evans, R. 2001. Applying machine learning toward an automatic classification of it. *Literary and linguistic computing*, 16(1):45–58.

Foraker, S. and B. McElree. 2007. The role of prominence in pronoun resolution: Active versus passive representations. *Journal of Memory and Language*, 56(3):357–383.

Ge, N., J. Hale, and E. Charniak. 1998. A statistical approach to anaphora resolution. In *Sixth Workshop on Very Large Corpora.*

Kennedy, A., R. Hill, and J. Pynte. 2003. The Dundee corpus. In *Proceedings of the 12th European Conference on Eye Movement (ECEM-2003).*

Lappin, S. and H. J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Li, Y., P. Musilek, M. Reformat, and L. Wyard-Scott. 2009. Identification of pleonastic it using the web. *J. Artif. Int. Res.*, 34(1):339–389, mar.

Liu, R., R. Mao, A. T. Luu, and E. Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review.*

Meng, Y. and A. Rumshisky. 2018. Triad-based neural network for coreference resolution. In E. M. Bender, L. Derczynski, and P. Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 35–43, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Mitkov, R. 1996. Towards a more efficient use of PC-based MT in education. In *Proceedings of Translating and the Computer 18*, London, UK, November 14-15. Aslib.

Mitkov, R. 1998. Robust pronoun resolution with limited knowledge. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 869–875, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Mitkov, R. 2002. *Anaphora Resolution.* Longman.

Mitkov, R., editor. 2003. *The Oxford Handbook of Computational Linguistics.* Oxford University Press.

Mitkov, R. 2006. Fully automatic anaphora resolution for English and Bulgarian. In S. K. M. Slavcheva, M. and G. Angelova, editors, *Readings in multilinguality.* Institute for parallel processing, Bulgarian Academy of Sciences, pages 78–86.

Mitkov, R. 2019. Computer vs. human intelligence. Keynote speech at the Refinitiv conference, City of London.

Mitkov, R., editor. 2022. *The Oxford Handbook of Computational Linguistics.* Oxford University Press, 2nd substantially revised edition.

Mitkov, R. and C. Barbu. 2000. Mutual enhancement of performance: bilingual pronoun resolution for English and French. In *DAARRC2000 - Discourse, Anaphora and Reference Resolution Conference.*

Mitkov, R., L. Belguith, and M. Stys. 1998. Multilingual robust anaphora resolution. In N. Ide and A. Voutilainen, editors, *Proceedings of the Third Conference on Empirical Methods for Natural Language Processing*, pages 7–16, Palacio de Exposiciones y Congresos, Granada, Spain, June. Association for Computational Linguistics.

Mitkov, R. and L. H. Belguith. 1998. Robust pronoun resolution with limited knowledge: a high success rate approach for English and Arabic. In *6th Iberoamerican Conference on Artificial Intelligence (IBERAMIA'98)*.

Mitkov, R., R. Evans, and C. Orasan. 2002. A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 168–186, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mitkov, R. and M. Stys. 1997. Robust reference resolution with limited knowledge: high precision genre-specific approach for English and Polish. In *Proceedings of the International Conference "Recent Advances in Natural Language Proceeding" (RANLP'97)*.

Nitoń, B., P. Morawiecki, and M. Ogrodniczuk. 2018. Deep neural networks for coreference resolution for Polish. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Orasan, C. and R. Evans. 2001. Learning to identify animate references. In *Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning (ConLL)*.

Peters, M., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 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)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Plu, J., R. Prokofyev, A. Tonon, P. Cudré-Mauroux, D. E. Difallah, R. Troncy, and G. Rizzo. 2018. Sanaphor++: Combining deep neural networks with semantics for coreference resolution. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Rayner, K. and S. A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201.

Rayner, K., A. Pollatsek, J. Ashby, and C. Clifton Jr. 2012. Psychology of reading.

Rohanian, O., S. Taslimipoor, V. Yaneva, and L. A. Ha. 2017. Using gaze data to predict multiword expressions. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 601–609, Varna, Bulgaria, September. INCOMA Ltd.

Rotsztejn, J. 2018. *Learning from cognitive features to support natural language processing tasks.* Master's thesis, ETH Zurich.

Stuckardt, R. 2002. Machine-learning-based vs. manually designed approaches to anaphor resolution: the best of two worlds. In *Proc.4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2002)*.

Stuckardt, R. 2003. Coreference-based summarization and question answering: a case for high precision anaphor resolution.

In *Proc. 2003 Int. Symp. Reference Resolution and Its Application to QA and TS(ARQAS)*.

Stuckardt, R. 2005. A machine learning approach to preference strategies for anaphor resolution. In *Anaphora Processing Linguistic, Cognitive and Computational Modeling* edited by Branco, A., McEnery, A., Mitkov, R. John Benjamins, Amsterdam/Philadelphia.

Tanev, H. and R. Mitkov. 2002. Shallow language processing architecture for Bulgarian. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Tapanainen, P. and T. Jarvinen. 1997. A non-projective dependency parser. In *Fifth Conference on Applied Natural Language Processing*, pages 64–71, Washington, DC, USA, March. Association for Computational Linguistics.

Trinh, T. H. and Q. V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

Vadász, N. 2023. Resolving Hungarian anaphora with ChatGPT. In K. Ekštein, F. Pártl, and M. Konopík, editors, *Text, Speech, and Dialogue*, pages 45–57, Cham. Springer Nature Switzerland.

Wiseman, S., A. M. Rush, and S. M. Shieber. 2016. Learning global features for coreference resolution. In K. Knight, A. Nenkova, and O. Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, San Diego, California, June. Association for Computational Linguistics.

Yaneva, V., L. A. Ha, R. Evans, and R. Mitkov. 2018. Classifying referential and non-referential it using gaze. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4896–4901, Brussels, Belgium, October-November. Association for Computational Linguistics.

Yang, X., E. Peynetti, V. Meerman, and C. Tanner. 2022. What GPT knows about who is who. In S. Tafreshi, J. Sedoc, A. Rogers, A. Drozd, A. Rumshisky, and A. Akula, editors, *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 75–81, Dublin, Ireland, May. Association for Computational Linguistics.