Classifying Referential and Non-referential *It* Using Gaze

Victoria Yaneva, Le An Ha, Richard Evans, and Ruslan Mitkov
Research Institute in Information and Language Processing,
University of Wolverhampton, UK
{v.yaneva, ha.l.a, r.j.evans, r.mitkov}@wlv.ac.uk

Abstract

When processing a text, humans and machines must disambiguate between different uses of the pronoun *it*, including non-referential, nominal anaphoric or clause anaphoric ones. In this paper, we use eye-tracking data to learn how humans perform this disambiguation. We use this knowledge to improve the automatic classification of *it*. We show that by using gaze data and a POS-tagger we are able to significantly outperform a common baseline and classify between three categories of *it* with an accuracy comparable to that of linguistic-based approaches. In addition, the discriminatory power of specific gaze features informs the way humans process the pronoun, which, to the best of our knowledge, has not been explored using data from a natural reading task.

1 Introduction

Anaphora resolution is both one of the most important and one of the least developed tasks in Natural Language Processing (Lee et al., 2016). A particularly difficult case for anaphora resolution systems is the pronoun *it*, as it may refer to a specific noun phrase or an entire clause, or it may even refer to nothing at all, as in sentences 1 - 3 below[1].

1. *"I couldnt say exactly, sir, but it wasnt teatime by a long way."* (Pleonastic *it* (non-referential)).
2. *Now, as to this quarrel. When was the first time you heard of it?* (Nominal anaphoric)
3. *You have been with your mistress many years, is it not so?* (Clause anaphoric[2].)

[1]Extracted from the GECO corpus (Cop et al., 2016)

[2]Some authors also distinguish other, less-common types of the pronoun it such as proaction, cataphoric, discourse topic, and idiomatic, among others (Evans, 2001).

This phenomenon is not specific only to English; pronouns that can be used both referentially and non-referentially exist in a variety of language groups such as the pronoun *'het'* in Dutch, *'det'* in Danish, *'ello'* in Spanish, *'il'* in French, etc.

In NLP, there has been active research in the area of automatic classification of *it* during the past four decades but the issue is far from being solved (Section 2). According to corpus statistics, the pronoun it is by far the most frequently used pronoun in the English language (Li et al., 2009), and as recently as 2016, incorrect classification of different cases of *it* and their antecedents is highlighted as one of the major reasons for the failure of question-answering systems (Lee et al., 2016). In the state of the art, the best approaches to classifying *it* achieve between 2% and 15% improvement over a majority baseline when assigning examples of the pronoun to more than two classes.

In contrast with the extensive research in NLP, very little is known about the way humans approach the disambiguation of *it*. To the best of our knowledge, Foraker and McElree (2007) is the only study on this subject that uses online measures of reading. It proves empirically that the pronoun *it* is resolved more slowly and less accurately than gendered pronouns due to its ambiguity. So far there have been no studies investigating the subject using natural reading data as opposed to artificially created controlled sentences.

In this paper we approach these two problems as one and hypothesise that obtaining information on the way humans disambiguate the pronoun can improve its automatic classification.

We propose a new method for classifying the pronoun *it* that does not rely on linguistic processing. Instead, the model leverages knowledge about the way in which humans disambiguate the pronoun based on eye tracking data. We

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4896–4901
Brussels, Belgium, October 31 - November 4, 2018.

show that by using gaze data and a POS tagger we are able to achieve classification accuracy of three types of *it* that is comparable to the performance of linguistic-based approaches and that outperforms a common baseline with a statistically significant difference. In addition, examining the discriminatory power of specific gaze features provides valuable information about the human processing of the pronoun. We make our data, code and annotation available at https://github.com/victoria-ianeva/It-Classification. The GECO eye-tracking corpus is available at http://expsy.ugent.be/downloads/geco/.

2 Related Work

Gaze data in NLP While eye-movement data has been traditionally used to gain understanding of the cognitive processing of text, it was recently applied to a number of technical tasks such as part-of-speech tagging (Barrett et al., 2016), detection of multi-word expressions (Rohanian et al., 2017; Yaneva et al., 2017), sentence compression (Klerke et al., 2016), complex-word identification (Štajner et al., 2017), and sentiment analysis (Rotsztejn, 2018). Eye movements were also shown to carry valuable information about the reader and were used to detect specific conditions affecting reading such as autism (Yaneva et al., 2018, 2016) and dyslexia (Rello and Ballesteros, 2015). The motivation behind these approaches is two-fold. First, eye tracking is already making its way into everyday use with interfaces and devices that feature eye-tracking navigation (e.g. Windows Eye Control[3]). Second, linguistic annotation by gaze is faster than traditional annotation techniques, does not require trained annotators, and provides a language-independent approach that can be applied to under-resourced languages (Barrett et al., 2016). This is particularly interesting for the case of non-referential pronouns, as the phenomenon exists in different languages.

Classification of *it* The majority of the machine learning approaches to classifying the pronoun *it* in different languages are based on linguistic features capturing token, syntactic and semantic context. Different papers report varying majority baseline metrics (between 50% and 75%) depending on the annotated corpora, and an improvement over the majority baselines of between 2% and 15% for classification of more than two classes of *it* (Loáiciga et al., 2017; Uryupina et al., 2016; Lee et al., 2016; Müller, 2006; Boyd et al., 2005; Hoste et al., 2007; Evans, 2001). For example, Loáiciga et al. (2017) train a bidirectional recurrent neural network (RNN) to classify three classes of *it* in the ParCor corpus (Guillou et al., 2014) and compare its performance to a feature-based maximum entropy classifier. The RNN achieves accuracy of 62% compared to a majority baseline of 54% and is significantly outperformed by the linguistic-feature classifier which obtained 68.7% accuracy. Lee et al. (2016) compare several statistical models and report 75% accuracy for four-class classification over a majority baseline of around 62% using linguistic features and a stochastic adaptive gradient algorithm. They also report that experimenting with word embeddings did not lead to more accurate classification. So far, approaches using linguistic features still represent the state of the art in the classification of *it*.

3 Data

Corpus: The eye-tracking data was extracted from the GECO corpus (see Cop et al. (2016) for full corpus specifications) which is the largest and most recent eye-tracking corpus for English at present. The text of the corpus is a novel by Agatha Christie entitled "The Mysterious Affair at Styles", the English version of which contains 54,364 tokens and 5,012 unique types. The entire novel was read by 14 native English undergraduates from the University of Southampton using an eye-tracker with a sampling rate of 1 kHz.

Annotation scheme: A total of 1,052 instances of *it* were found in the corpus[4]. Each of the instances of *it* was annotated by two annotators and assigned to one of three categories: *Pleonastic, Nominal anaphoric* and *Clause anaphoric*, following the scheme used by Lee et al. (2016). The annotators were free to view as much of the previous text as necessary to decide on a label. The inter-annotator agreement for the three categories was $\kappa = 0.636$, $p < 0.0005$, indicating substantial agreement between the annotators. This number corresponds to a percentage agreement of 77.47% for the three categories and is comparable to the

[3]https://support.microsoft.com/en-gb/help/4043921/windows-10-get-started-eye-control

[4]This number does not include the possessive pronoun *its*. There are also several tokenisation errors in the GECO corpus (e.g. "it?..ah," and "it...the" misidentified as single tokens.). These cases were excluded.

	Annotat. 1	Annotat. 2	Final
Pleonastic	339 (33%)	406 (38%)	272 (33%)
Nom. anaph.	492 (46%)	527 (50%)	453 (56%)
Clause anaph.	221 (21%)	119 (11%)	89 (11%)

Table 1: Annotation categories

	Prev.	"It"	Next	It + Next
Early	61.1	60.3	61.7	61
Medium	60.9	60.5	60.6	61.2
Late	58.9	60.2	61	61.4

Table 2: Weighted F1 scores for an ablation study for different gaze feature groups over the Previous and Next word baseline (60.4)

81% agreement reported in Lee et al. (2016). We perceived adjudication between cases of disagreement (237 instances) to be extremely arbitrary, so those cases were excluded rather than resolved. Examples of such arbitrary cases include:

- *"Sit down here on the grass, do. It's ever so much nicer."* (nominal vs. clause anaphoric)
- *It's a jolly good life taking it all round...if it weren't for that fellow Alfred Inglethorp!"* (pleonastic vs. clause anaphoric)

The distribution of each class of the retained data by annotator is presented in Table 1. We make the full annotations of both annotators available.

4 Experiments

Overview In order to test the extent to which gaze data can help the classification of different cases of *it*, we trained and compared three separate classifiers. The first classifier is based on gaze features, the second one is based on linguistic features and finally, we trained a combined classifier using both gaze and linguistic features. We compared the performance of these classifiers to a majority baseline of 55.7 and to another baseline obtained by using the tokens surrounding the pronoun (previous and next word) as features (60.4). We also experimented with adding word embeddings[6] for the surrounding tokens as features. While the full exploration of word embeddings for the classification of *it* remains outside of the scope of this work, it would be interesting to explore whether the embeddings add value to the models by encoding information that was not otherwise captured.

[5]Except L4 and L3

[6]300-dimensional vectors from Google News obtained through word2vec: https://code.google.com/archive/p/word2vec/

		Ling	Gaze	Comb
EARLY	First_Run_Fixation_Count		*	
	First_Run_Fixation_%			*
	First_Fixation_Duration			
	First_Fixation_Visited_Count			†*
	First_Fix_Progressive			†*
MEDIUM	Second_Run_Fixation_Count			
	Second_Run_Fixation_%			*
	Second_Fixation_Duration		*	†*
	Second_Fixation_Run		*	*
	Gaze_Duration			†
LATE	Third_Run_Fixation_Count			
	Third_Run_Fixation_%			†
	Third_Fixation_Duration			†*
	Third_Fixation_Run		*	
	Last_Fixation_Duration		*	†*
	Last_Fixation_Run			
	Go_Past_Time			†
	Selective_Go_Past_Time			†
	Fixation_Count			†
	Fixation_%		*	†
	Total_Reading_Time			
	Total_Reading_Time_%		*	*
	Trial_Fixation_Count			*
	Trial_Total_Reading_Time			
	Spillover			
	Skip		*	
LINGUISTIC	Word_position	+		+
	# Preceding_NPs_in_sentence			
	# Preceding_NPs_in_paragraph			
	# Following_NPs_in_sentence			
	# NPs_in_the_sentence			
	# NPs_in_the_paragraph			+
	# Following_adject._in_sentence			
	Previous_verb	+		
	Following_adjective	+		+
	Following _verb	+		+
	POS in posit. L4, L3, L2, L1	+	+[5]	+
	POS in posit. R4, R3, R2, R1	+	+	+
	# Following_complementisers	+		+
	An_adjective_before_the_next_NP	+		+
	Words_until_next_complementiser	+		+
	Words_until_next_infinitive	+		+
	Words_until_next_preposition			+
	Words_until_next_ing_verb	+		+
	A_compl._before_the_next_NP			
	Immediately_preceding_preposit.			
BASIC	Previous_word	+	+	+
	Next_word	+	+	+
	Word_length		+	+
	Punctuation			+

Table 3: List of features and their inclusion in the different models. + refers to linguistic data, * to added values for the *It + Next* region, and †to gaze features for the previous word region. The features that do not have marks in the last three columns were not retained in any of the three best models.

Gaze features We use the gaze features as provided in GECO and we average the data from all 14 readers per token. We extract gaze data for each case of *it*, as well as for the preceding and following word. The full list of gaze features used in the

experiments can be seen in Table 3.

Different eye-tracking measures (usually divided into *early* and *late*) are indicative of different aspects of cognitive processing. Early gaze measures such as *First_Fixation_Duration* give information about the early stages of lexical access and syntactic processing, while late measures such as *Total_Reading_Time* or *Number_of_Fixations* give information about processes such as textual integration and disambiguation (see Rayner et al. (2012) for a review). In Table 3, the distinction between *Early, Medium* and *Late* gaze features is mainly based on the run during which the fixations were made (i.e. whether the eyes were passing through the text for the first, second or third time). For each run we report count measures, percentage measures (as part of the trial) and duration measures (in milliseconds). Additional late features reported include *Last_Fixation_Duration* and *Last_Fixation_Run* (the run during which the last fixation in a given region occurred), *Total_Reading_Time* (in msec and %), and *Trial_Fixation_Count* (the overall number of fixations within the trial). *Go_Past_Time* refers to the summation of all fixation durations on the current word during the first pass. *Spillover* refers to the duration of the first fixation made on the next word after leaving the current word in the first run. Finally, a word is considered skipped if no fixation occurred during the first run (*Skip*). A complete legend explaining each feature can be found within the corpus metadata.

An ablation study on the contributions of individual groups of gaze features towards the classification of *it* is presented in Table 2.

Linguistic features We implemented a set of features originally proposed by Evans (2001) and subsequently used extensively in the studies presented in Section 2. In terms of features and categories of *it*, the study by Evans (2001) is the most fine-grained one we could find, classifying 7 categories of *it* with 69% accuracy. The set of features from Evans (2001) is presented in Table 3. These features synthesize information based on corpus studies of the pronoun *it* and thus aim to capture positional, part-of-speech and proximity information, as well as specific patterns of usage. For example, Evans (2001) notes that pleonastic pronouns rarely appear immediately after a prepositional word and that complementisers or adjectives often follow pleonastic instances. Another pattern that distinguishes the pleonastic use of *it* is associated with certain sequences of elements such as 'adjective + noun phrase' and 'complementiser + noun phrase' (Evans, 2001). Therefore, the linguistic features proposed by Evans (2001) and used in our experiments make possible the utilization of corpus-based knowledge for the automatic classification of *it*.

	P	R	F1
Baselines			
Majority baseline			55.7
Previous + Next word	62.1	63.9	60.4
Embeddings			
Prev. + Next Embed.	63.1	64.4	62.1
Linguistic models			
Full feature set	63.4	66	63.2
Best linguistic	**66.7**	**68.8**	**66.1***
Best linguistic + Embed.	**66.9**	**68.8**	**67.2***
Gaze-based models			
Basic + POS	63.3	64.5	62.2
Select. Gaze + Basic	65.8	66.8	64.2
Select. Gaze + Basic + POS	**66.6**	**67.9**	**65.6***
S. Gaze + Bas. + POS + Embed.	**66.3**	**68.8**	**66.7***
Combined model			
Best Gaze + Ling	**71**	**71.5**	**68.8***
Best Gaze + Ling. + Embed.	67.5	69.3	671*

Table 4: Precision, Recall and Weighted F1 for the various classifiers. The * symbol marks statistical significance compared to the baseline model of Previous + Next Word (60.4).

Classification We use simple logistic regression as implemented in WEKA (Hall et al., 2009) with 10-fold cross validation and a random seed parameter 20. Since logistic regression is an interpretable method, we are able to assess the performance of individual features and gain insight into the psycholinguistic processing of the pronoun.

We experimented with gaze features for the individual words but as gaze data is inherently very noisy, we found that smoothing the features by adding the ones that correspond to the pronoun and the next word stabilized the results. Adding the gaze features for the previous word significantly reduced the performance but using them separately in a model with the added *It + Next* features maximised our results. For the role of individual features in the models see Table 3.

In order to account for class imbalance we compute and report a weighted F1 score, as opposed to the harmonic mean between precision and recall. First, the F1 for each class is weighted by multi-

plying it by the number of instances in the class. Then the F1 scores for all classes are summed up and divided by the total number of instances. The resulting weighted F1 score is lower than the traditionally reported mean F1 score, but it represents the effects of class imbalance more accurately.

5 Results and Discussion

The results from the classification experiments presented in Table 4 have implications for language processing by both humans and machines.

From the NLP perspective, the potential of gaze data to not only improve but also, to a certain extent, substitute text processing approaches is an exciting new frontier. Our results show that the gaze-based classifier performs on par with the one using linguistic features and both of them perform significantly better[7] than the baseline of 60.4. An improvement of 13% over the majority baseline is achieved when combining the two but this difference is not a significant improvement over the individual best classifiers. A possible reason for this is that the gaze data and linguistic features encode similar information about the disambiguation of *it* and adding them together leads to overlap instead of complementation. In all three classifiers, the clause-anaphoric class was consistently predicted with lowest accuracy (Table 5), which is not surprising given that it only accounts for 11% of the retained data. In line with the observation of Lee et al. (2016) (Section 2), the embeddings do not show a stable contribution. In our case, this is likely related to the small amount of data, to which we add 300 dimensions per word.

Overall, the improvement achieved by the classifiers is comparable with the current state-of-the-art (Section 2). It is important to note that this is the first study to use text from the domain of literature and that this may have influenced the extraction of the linguistic features. At the same time, literature can be regarded as a more challenging domain than the declarative texts used in previous research, owing to the creative use of language.

From a psycholinguistic perspective, we provide evidence that, indeed, the three classes of *it* are processed differently. We observe that medium and late gaze features related to disambiguation are more discriminative than the early ones. For example, measures such as first fixation duration were not included in any of the models, while revisits as late as the third run (the third time the eyes pass over the region of interest) occurred in these regions and provided a strong signal. Particularly useful features were the durations of the second and last fixations, as well as the information about the run (pass) during which they occur.

NomAnaph	ClauseAnaph	Pleon	
395	2	56	NomAnaph
53	11	25	ClauseAnaph
93	3	176	Pleon

Table 5: Confusion matrix for the best combined model (Weighted F1 = 68.8)

The significance of these features in the best classifiers somewhat contradicts the ablation study presented in Table 2. According to that table, early processing features for the preceding and next words are expected to outperform the late ones. A possible explanation for this are predictability and spillover effects, as the pronoun *it* is both highly predictable and easy to skip, because of its high frequency and shortness. Indeed, the gaze features from the it-region itself are not as useful as the ones from the surrounding words.

The results from this study showed that: i) gaze features encode information about the way humans disambiguate the pronoun *it*, ii) that this information partially overlaps with the information carried by linguistic features, and that iii) gaze can be used for automatic classification of the pronoun with accuracy comparable to that of linguistic-based approaches. In our future work we will attempt to identify specific patterns of cognitive processing for the individual classes, as well as explore factors related to the readers.

6 Conclusion

We presented the first study on the use of gaze data for disambiguating categories of *it*, exploring a wide range of gaze and linguistic features. The model based on gaze features and part-of-speech information achieved accuracy similar to that of the linguistic-based model and state-of-the-art systems, without the need for text processing. Late gaze features emerged as the most discriminative ones, with disambiguation effort indicators as late as third pass revisits.

[7] Gaze + Basic + POS: $p = 0.029$, 95% CI (0.509; 9.858) ; Best Linguistic: $p = 0.0017$, 95% CI (1.01; 10.34); Best Gaze + Ling: $p = 0.0004$, 95% CI (3.75; 12.99). The CI indicate the difference in %

References

Maria Barrett, Joachim Bingel, Frank Keller, and Anders 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)*, volume 2, pages 579–584.

Adriane Boyd, Whitney Gegg-Harrison, and Donna Byron. 2005. Identifying non-referential it: a machine learning approach incorporating linguistically motivated patterns. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 40–47. Association for Computational Linguistics.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2016. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, pages 1–14.

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

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

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. 2014. Parcor 1.0: A parallel pronoun-coreference corpus to support statistical mt. In *9th International Conference on Language Resources and Evaluation (LREC), MAY 26-31, 2014, Reykjavik, ICELAND*, pages 3191–3198. European Language Resources Association.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Veronique Hoste, Iris Hendrickx, and Walter Daelemans. 2007. Disambiguation of the neuter pronoun and its effect on pronominal coreference resolution. In *International Conference on Text, Speech and Dialogue*, pages 48–55. Springer.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. *arXiv preprint arXiv:1604.03357*.

Timothy Lee, Alex Lutz, and Jinho D Choi. 2016. Qait: Classifying non-referential it for question answer pairs. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 132–137.

Yifan Li, Petr Musilek, Marek Reformat, and Loren Wyard-Scott. 2009. Identification of pleonastic it using the web. *Journal of Artificial Intelligence Research*, 34:339–389.

Sharid Loáiciga, Liane Guillou, and Christian Hardmeier. 2017. What is it? disambiguating the different readings of the pronoun it. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1325–1331.

Christoph Müller. 2006. Automatic detection of non-referential it in spoken multi-party dialog. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr. 2012. *Psychology of reading*. Psychology Press.

Luz Rello and Miguel Ballesteros. 2015. Detecting readers with dyslexia using machine learning with eye tracking measures. In *Proceedings of the 12th Web for All Conference*, page 16. ACM.

Omid Rohanian, Shiva Taslimipoor, Victoria Yaneva, and Le An 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.

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

Sanja Štajner, Victoria Yaneva, Ruslan Mitkov, and Simone Paolo Ponzetto. 2017. Effects of lexical properties on viewing time per word in autistic and neurotypical readers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 271–281.

Olga Uryupina, Mijail Kabadjov, and Massimo Poesio. 2016. Detecting non-reference and non-anaphoricity. In *Anaphora Resolution*, pages 369–392. Springer.

Victoria Yaneva, Le An Ha, Sukru Eraslan, Yeliz Yesilada, and Ruslan Mitkov. 2018. Detecting autism based on eye-tracking data from web searching tasks. In *Proceedings of the Internet of Accessible Things*. ACM.

Victoria Yaneva, Shiva Taslimipoor, Omid Rohanian, and Le An Ha. 2017. Cognitive processing of multiword expressions in native and non-native speakers of english: Evidence from gaze data. In *International Conference on Computational and Corpus-Based Phraseology*, pages 363–379. Springer.

Victoria Yaneva, Irina P Temnikova, and Ruslan Mitkov. 2016. A corpus of text data and gaze fixations from autistic and non-autistic adults. In *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference (LREC)*.