Leveraging Annotators' Gaze Behaviour for Coreference Resolution

Joe Cheri Ross, Abhijit Mishra, Pushpak Bhattacharyya
Department of Computer Science & Engineering
Indian Institute of Technology Bombay, Mumbai
{joe,abhijitmishra,pb}@cse.iitb.ac.in

Abstract

This paper aims at utilizing cognitive information obtained from the eye movements behavior of annotators for automatic coreference resolution. We first record eye-movement behavior of multiple annotators resolving coreferences in 22 documents selected from MUC dataset. By inspecting the gaze-regression profiles of our participants, we observe how regressive saccades account for selection of potential antecedents for a certain anaphoric mention. Based on this observation, we then propose a heuristic to utilize gaze data to prune mention pairs in mention-pair model, a popular paradigm for automatic coreference resolution. Consistent improvement in accuracy across several classifiers is observed with our heuristic, demonstrating why cognitive data can be useful for a difficult task like coreference resolution.

1 Introduction

Coreference resolution deals with identifying the expressions in a discourse referring to the same entity. It is crucial to many information retrieval tasks (Elango, 2005). One of its main objectives of is to resolve the noun phrases to the entities they refer to. Though there exist many rule based (Kennedy and Boguraev, 1996; Mitkov, 1998; Raghunathan et al., 2010) and machine learning based (Soon et al., 2001; Ng and Cardie, 2002; Rahman and Ng, 2011) approaches to coreference resolution, they are way behind imitating the human process of coreference resolution. Comparing the performance of different existing systems on a standard dataset, *Ontonotes*, released for CoNLL-2012 shared task (Pradhan et al., 2012), it is quite evident that the recent systems do not have much improvement in accuracy over the earlier systems (Björkelund and Farkas, 2012; Durrett and Klein, 2013; Björkelund and Kuhn, 2014; Martschat et al., 2015; Clark and Manning, 2015).

This paper attempts to gain insight into the cognitive aspects of coreference resolution to improve mention-pair model, a well-known supervised coreference resolution paradigm. For this we employ eye-tracking technology that has been quite effective in the field of psycholinguistics to study language comprehension (Rayner and Sereno, 1994), lexical (Rayner and Duffy, 1986) and syntactic processing(von der Malsburg and Vasishth, 2011). Recently, eye-tracking studies have been conducted for various language processing tasks like Sentiment Analysis, Translation and Word Sense Disambiguation. Joshi et al. (2014) develop a method to measure the sentiment annotation complexity using cognitive evidence from eye-tracking. Mishra et al. (2013) measure complexity in text to be translated based on gaze input of translators which is used to label training data. Joshi et al. (2013) propose a studied the cognitive aspects if Word Sense Disambiguation (WSD) through eye-tracking.

Eye-tracking studies have also been conducted for the task of coreference resolution. Cunnings et al. (2014) check for whether the syntax or discourse representation has better role in pronoun interpretation. Arnold et al. (2000) examine the effect of gender information and accessibility to pronoun interpretation. Vonk (1984) studies the fixation patterns on pronoun and associated verb phrases to explain comprehension of pronouns.

We perform yet another eye-tracking study to understand certain facets of human process involved in coreference resolution that eventually can help automatic coreference resolution. Our participants are given a set of documents to perform coreference annotation and the eye movements during the exercise are recorded. Eye-movement patterns are characterized by two basic attributes: (1) Fixations, corresponding to a longer stay of gaze on a visual object (like charac-

Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning, pages 22–26,
Berlin, Germany, August 11, 2016.

ters, words *etc.* in text) (2) Saccades, corresponding to the transition of eyes between two fixations. Moreover, a saccade is called a *Regressive Saccade* or simply, *Regression* if it represents a phenomenon of going back to a pre-visited segment. While analyzing these attributes in our dataset, we observe a correlation between the *Total Regression Count* and the complexity of a mention being resolved. Additionally, *Mention Regression Count*, *i.e.,* the count of a previous mention getting visited while resolving for an anaphoric mention, proves to be a measure of relevance of that particular mention as antecedent to the anaphoric mention.

Following the insights, we try to enrich mention-pair model, a popular paradigm in automatic coreference resolution by performing mention pair pruning prior to classification using mention regression data.

2 Creation of Eye-movement Database

We prepared a set of 22 short documents, each having less than 10 sentences. These were selected from the MUC-6 dataset[1]. Discourse size is restricted in order to make the task simpler for the participants and to reduce eye movements error caused due to scrolling.

The documents are annotated by 14 participants. Out of them, 12 of them are graduate/postgraduate students with science and engineering background in the age group of 20-30 years, with English as the primary language of academic instruction. The rest 2 are expert linguists and they belong to the age group of 47-50. To ensure that they possess good English proficiency, a small English comprehension test is carried out before the start of the experiment. Once they clear the comprehension test, they are given a set of instructions beforehand and are advised to seek clarifications before they proceed further. The instructions mention the nature of the task, annotation input method, and necessity of head movement minimization during the experiment.

The task given to the participants is to read one document at a time, and assign ids to mentions that are already marked in the document. Each id corresponding to a certain mention has to be unique, such that all the coreferent mentions in a single coreference chain are assigned with the same id. During the annotation, eye movements data of the participants (in terms of fixations, saccades and pupil-size) are tracked using an SR-Research Eyelink-1000 Plus eye-tracker (monocular mode with sampling rate of 500 Hz). The eye-tracking device is calibrated at the start of each reading session. Participants are allowed to take breaks between two reading sessions, to prevent fatigue over time.

We observe that the average annotation accuracy in terms of CoNLL-score ranges between **70.75%-86.81%**. Annotation error, we believe, could be attributed to: (a) Lack of patience/attention while reading, (b) Issues related to text comprehension and understanding, and (c) Confusion/indecisiveness caused due to lack of context. The dataset is freely available for academic use[2].

3 Analysis of *Eye-regression* Profiles

The cognitive activity involved in resolving coreferences is reflected in the eye movements of the participants, especially in the movements to the previously visited words/phrases in the document, termed as *regressive saccades* or simply, *regressions*. Regression count refers to the number of times the participant has revisited a candidate antecedent mention while resolving a particular anaphoric mention. This is extracted from the eye movement events between the first gaze of the anaphoric mention under consideration and the annotation event of this mention (when participants annotate the mention with a coreferent id).

Figure 1 shows the mention position (for a given mention id) in terms of the order of the mention in the document against count of regression going out from each mention to the previous mentions. The regression count for a particular mention is averaged over all the participants. As we see, average regression count tends to increase with increase in mention id, except for some mentions which may not have required visiting to the previous mentions for resolving them. The complexity of the content in MUC-6 dataset makes the spread of the regression counts dispersed. We also observe that, towards the end of the document, participants tend to regress more to the earlier sections because of limited working memory (Calvo, 2001). This increases the number of regressions performed from mentions appearing towards the

[1] `http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html`

[2] http://www.cfilt.iitb.ac.in/cognitive-nlp/

end of the document.

It is worth noting that intra-sentential mentions that have antecedents within the same sentence (as in '*Prime Minister **Brian Mulroney** and **his** cabinet have been briefed today*') do not generally elicit regressions. We believe, intra-sentential resolutions are connected to processing of syntactic constraints in an organized manner, as explained by the binding theory (Chomsky, 1982). Though the number of intra-sentential mentions in our dataset is low, it is evident from figure 1, that they do not account for many regressions.

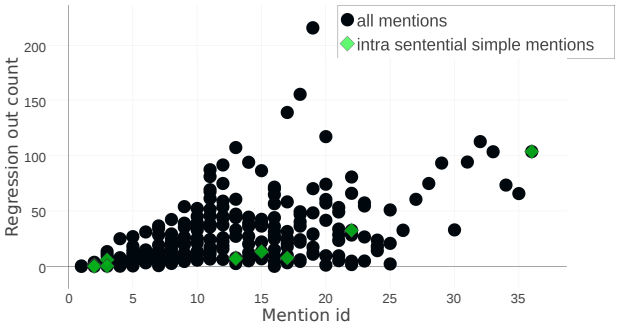

Figure 1: MUC-6 dataset: Mention id Vs Regression count

This above analysis on regression counts supports our hypothesis that the mentions that are regressed to more frequently have a better say in resolving an anaphoric mention.

4 Leveraging Cognitive Information Automatic Coreference Resolution

We experiment with a supervised system following a mention-pair model (Soon et al., 2001)-injecting the eye-movement information into it. Mention-pair model classifies mention pairs formed between mentions in a document as coreferent or not, followed by clustering, forming clusters of coreferent mentions. Eye tracking information is utilized in the process of mention pair pruning prior to mention pair classification.

4.1 For Mention-pair Pruning

Given an anaphoric mention, the probability of each previous mention being selected as antecedent is computed as follows. Transitions done by a participant to potential antecedent mentions, while resolving an anaphoric mention, are first obtained from the regression profile. From this, we filter out the regressions to a candidate antecedent mention that happen between two events- (a) first fixation lands on the anaphoric mention and (b) the anaphoric mention gets annotated with an id.

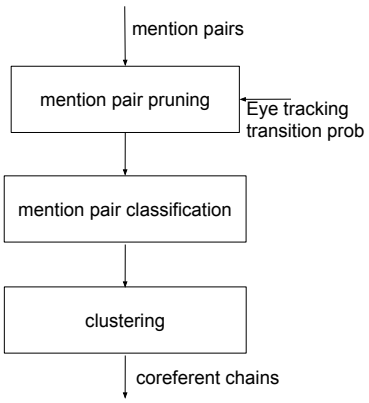

Figure 2: Mention-pair pruning

These regression counts from all the participants are aggregated to compute the transition probability values, as follows:

$$P_{m_i,m_j} = \frac{count(transitions\ m_j \rightarrow m_i)}{\sum_k count(transitions\ m_j \rightarrow m_k)} \quad (1)$$

In equation 1, P_{m_i,m_j} gives the transition probability value for an anaphoric mention m_j to a candidate antecedent mention m_i. count() computes the aggregated regression count over all participants. Denominator part computes for all candidate antecedents(k) of the anaphoric mention.

Transition probability thus computed for candidate mention pairs, are utilized prior to mention pair classification, filtering out irrelevant mention pairs. In the mention pair model, a mention pair(m_{ant}, m_{ana}) is formed between an anaphoric mention (m_{ana}) and a candidate antecedent mention (m_{ant}). For an anaphoric mention, the threshold probability value is computed from the number of potential candidate antecedents. $P_{thresh} = \frac{1}{\#candidate\ antecedents}$. Mentions pairs having probability less than P_{thresh} are pruned.

5 Experiments and Results

Eye movement data driven mention pair pruning, as discussed above, is experimented across different classifiers, *viz.*, Support Vector Machine (SVM), Naive Bayes, and Multi-layered Feed-Forward Neural Network (Neural Net). We use libsvm[3] for SVM implementation and Scikit-Learn[4] for Naive Bayes implementation. The neu-

[3]https://www.csie.ntu.edu.tw/ cjlin/libsvm/

[4]http://scikit-learn.org/

Experiments		MUC			B^3			CEAFe			CoNLL
		P	R	F	P	R	F	P	R	F	
SVM (RBF)	unpruned	61.13	68.96	64.81	57.72	75.39	65.38	47.33	58.23	52.22	**60.80**
	pruned	62.67	66.99	64.76	62.62	73.71	67.71	52.00	57.83	54.76	**62.41**
SVM (Linear)	unpruned	53.33	70.93	60.88	37.64	75.02	50.13	26.56	51.27	34.99	**48.67**
	pruned	54.71	71.42	61.96	39.63	75.07	51.88	29.44	53.14	37.89	**50.58**
Naive Bayes	unpruned	62.85	97.53	76.44	23.23	98.03	37.56	10.53	54.22	17.64	**43.88**
	pruned	62.90	96.05	76.02	25.06	96.64	39.80	13.50	58.64	21.94	**45.92**
Neural Net	unpruned	64.73	71.42	67.91	63.71	77.20	69.81	52.60	61.96	56.90	**64.87**
	pruned	66.35	70.93	68.57	66.55	76.15	71.03	55.76	62.01	58.72	**66.11**
Berkeley coref	unpruned	84.89	58.12	69.0	84.93	47.86	61.22	82.45	37.96	51.99	**60.73**
	pruned	86.86	58.62	70.0	87.15	47.64	61.6	82.7	39.26	53.25	**61.61**

Table 1: Results with different classifiers and Berkeley coreference system with and without pruning of candidate mention pairs (P,R,F)→ (Precision, R:Recall, F:F-measure), CoNLL:CoNLL Score

ral network classifier having an input layer, a hidden layer and an output layer is implemented using Keras[5]. For training, we consider a subset of English section of OntoNotes (v5.0) data (Pradhan et al., 2012) with 1634 documents. Testing is done with the 22 documents taken from MUC-6 dataset.

Since the main aspect of our work is mention pair pruning, we first check the mention pair pruning accuracy. We find that mention pair pruning has a precision of **87.24%**. Pruning errors may be attributed to increased number of regressions happening to mentions towards the end of the documents (refer section 3).

Performance of the system is evaluated using MUC, B^3 and CEAFe metrics. CoNLL score is computed as the average of F1s of all the mentioned metrics. Table 1 shows the results across different classifiers with and without mention pair pruning. Considering the CoNLL score, there is an improvement in performance across all classifiers. This improvement is contributed by the increase in precision , despite the fall in recall. Table 2 shows a few instances of non-coreferent antecedent-anaphora pairs which are correctly predicted as non-coreferent because of pruning.

Antecedent	**Anaphora**
here	*a treaty*
Paramount Communications Inc	*an after-tax gain of $1.2 billion*
Rogers Communications	*A Spokesman*

Table 2: Instances of precision errors corrected by pruning

Among all the classifiers neural network gives better accuracy, but the effective performance gain is higher with classifiers with lesser accuracy. Naive Bayes giving the least accuracy, gives the best accuracy improvement of 2.04% with mention-pair pruning. This gives the impression that systems with lower performance, are likely to benefit from the eye movement based heuristics.

The performance improvement of mention pair pruning is also verified with the state of the art Berkeley Coreference Resolution system (Durrett and Klein, 2013). The choice of the system was based on the code accessibility to make the modification required for mention pair pruning. Results of Berkeley system in table 1 shows that there is an improvement in CoNLL score , mainly contributed by the increase in precision.

6 Conclusion and Future Work

As far as we know, our work of utilizing cognitive information for the task of automatic coreference resolution is the first of it kind. By analyzing the eye-movement patterns of annotators, we observe a correlation between the complexity of resolving an anaphoric mention and eye-regression count associated with the preceding mentions. We also observe that gaze transition probability derived from regression counts associated with a mention signify the candidacy of that mention as an antecedent. This helps us devise a heuristic to prune irrelevant mention pair candidates in a supervised coreference resolution approach. Our heuristic brings noticeable improvement in accuracy with different classifiers. The current work can be further enriched to utilize eye-gaze information for (a) meaningful feature extraction for mention pair classification and (b) proposing efficient clustering mechanism. We would also like to replace our current annotation setting with a non-intrusive reading setting (say, reading text on mobile devices with camera based eye-trackers), where explicit annotations need not be required.

[5]http://keras.io/

Acknowledgments

We thank for the support of CFILT lab at IIT Bombay and the annotators who helped us with coreference annotation experiment.

References

Jennifer E Arnold, Janet G Eisenband, Sarah Brown-Schmidt, and John C Trueswell. 2000. The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, 76(1):B13–B26.

Anders Björkelund and Richárd Farkas. 2012. Data-driven multilingual coreference resolution using resolver stacking. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 49–55. Association for Computational Linguistics.

Anders Björkelund and Jonas Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the Association for Computational Linguistics*.

Manuel G Calvo. 2001. Working memory and inferences: Evidence from eye fixations during reading. *Memory*, 9(4-6):365–381.

Noam Chomsky. 1982. *Some concepts and consequences of the theory of government and binding*, volume 6. MIT press.

Kevin Clark and Christopher D Manning. 2015. Entity-centric coreference resolution with model stacking. In *Association of Computational Linguistics (ACL)*.

Ian Cunnings, Clare Patterson, and Claudia Felser. 2014. Variable binding and coreference in sentence comprehension: evidence from eye movements. *Journal of Memory and Language*, 71(1):39–56.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *EMNLP*, pages 1971–1982.

Pradheep Elango. 2005. Coreference resolution: A survey. *University of Wisconsin, Madison, WI*.

Salil Joshi, Diptesh Kanojia, and Pushpak Bhattacharyya. 2013. More than meets the eye: Study of human cognition in sense annotation. In *HLT-NAACL*, pages 733–738.

Aditya Joshi, Abhijit Mishra, Nivvedan Senthamilselvan, and Pushpak Bhattacharyya. 2014. Measuring sentiment annotation complexity of text. In *ACL (2)*, pages 36–41.

Christopher Kennedy and Branimir Boguraev. 1996. Anaphora for everyone: pronominal anaphora resoluation without a parser. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 113–118. Association for Computational Linguistics.

Sebastian Martschat, Patrick Claus, and Michael Strube. 2015. Plug latent structures and play coreference resolution. *ACL-IJCNLP 2015*, page 61.

Abhijit Mishra, Pushpak Bhattacharyya, Michael Carl, and IBC CRITT. 2013. Automatically predicting sentence translation difficulty. In *ACL (2)*, pages 346–351.

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

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2011. Syntactic parsing for ranking-based coreference resolution. In *IJCNLP*, pages 465–473.

Keith Rayner and Susan 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.

Keith Rayner and Sara C Sereno. 1994. Eye movements in reading: Psycholinguistic studies. *Handbook of Psycholinguistics*.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Titus von der Malsburg and Shravan Vasishth. 2011. What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, 65(2):109–127.

W Vonk. 1984. Eye movements during comprehension of pronouns. *Advances in Psychology*, 22:203–212.