A Survey on Using Gaze Behaviour for Natural Language Processing

Sandeep Mathias\textsuperscript{1}, Diptesh Kanojia\textsuperscript{1,2}, Abhijit Mishra* and Pushpak Bhattacharyya\textsuperscript{1}
\textsuperscript{1}Indian Institute of Technology Bombay
\textsuperscript{2}IITB-Monash Research Academy
{sam, diptesh, pb}@cse.iitb.ac.in, abhijitmishra.530@gmail.com

\section*{Abstract}
Gaze behaviour has been used as a way to gather cognitive information for a number of years. In this paper, we discuss the use of gaze behaviour in solving different tasks in natural language processing (NLP). The collection of gaze behaviour is a costly task, both in terms of time and money. Hence, in this paper, we focus on research done to alleviate the need for recording gaze behaviour at run time. We also describe different eye tracking corpora in multiple languages, which are currently available and can be used in natural language processing. We conclude our paper by discussing applications in a domain - education - and how learning gaze behaviour can help in solving the tasks of complex word identification and automatic essay grading.

\section{Introduction}
Collecting psycholinguistic information from a reader has benefited multiple tasks in NLP, like named-entity recognition (NER) [Hollenstein and Zhang, 2019], text quality rating prediction [Mathias et al., 2018], sarcasm understandability [Mishra et al., 2016a], etc. Gaze behaviour, in particular, has been shown to correlate well with cognitive processing of text, via the eye-mind hypothesis, which states that “there is no appreciable lag between what is fixated and what is processed” [Just and Carpenter, 1980]. Therefore, while using gaze behaviour is helpful for solving NLP tasks, a massive challenge involved is how do we collect the gaze behaviour in the first place? In this paper, we describe research that uses gaze behaviour at run time to solve different NLP tasks.

While studying the gaze behaviour of a reader, we define the following terms. An \textbf{Interest Area} is the part of the screen that is of interest for analysis. In the case of NLP, it is mainly the words, although it could also be phrases, sentences, and paragraphs. A \textbf{Fixation} is an event where the eye is focused on the screen. Fixations take place when the eye is processing what it sees - in this case, the text that is being read. A \textbf{Saccade} is the movement of the eye from one fixation point to the next. There are 2 types of saccades - \textbf{Progressions} and \textbf{Regressions}. Progressions take place when there is a saccade from a fixation in the current interest area to a fixation in a later interest area. Regressions take place when there is a saccade from a fixation in the current interest area to a fixation in an earlier interest area. Figure 1 shows the output of an eye-tracker (the SR Research Eye Link 1000), for a reader reading a piece of text. The yellow boxes correspond to the interest areas, the cyan circles correspond to fixations, and the dark-yellow arrows correspond to saccades.

\section{Motivation for Learning Gaze Behaviour}
One of the earliest works that postulated the utility of gaze behaviour was a study on reading comprehension done by Just and Carpenter [1980]. They came up with the eye-mind hypothesis, which stated that “there is no appreciable lag between what is fixated by the eye and what is processed by the mind.” Many studies, done in the field of cognitive science, have found relationships between different aspects of gaze behaviour and corresponding aspects of text, such as relationships between fixations and word length [Rayner, 1998; Henderson and Ferreira, 1993], word predictability [Rayner, 1998], etc.

The gaze behaviour of a reader can therefore provide valuable psycholinguistic information for systems to help them solve tasks that require human intelligence. For example, using gaze behaviour, we can better quantify the difficulty of sentences for translation [Mishra et al., 2013], or verify if a reader understands the sarcasm in a piece of text [Mishra et al., 2016a], or even evaluate the quality of word embeddings [Søgaard, 2016]. Portable devices and applications (like Samsung Smart Scroll) generally provide a lower quality of gaze behaviour recording compared to research-grade devices. Our goal in this paper is to introduce the AI audience to ways in which we alleviate the need for readers to read texts at run time for the purpose of collecting gaze behaviour.

\section{Eye Tracking Corpora}
There are several publicly available corpora where readers solve different tasks in different languages\textsuperscript{1}. One of the earliest publicly available eye-tracking corpus is the Dundee Corpus [Kennedy et al., 2003]. The Dundee Corpus was created for both English and French, with the English version having

\footnote{\textsuperscript{*}currently employed at Apple Inc.}

\footnote{\textsuperscript{1}A number of those corpora can be downloaded from here.}
20 articles from The Independent read by 10 readers, and the French version having 20 articles from Le Monde read by 10 French speakers.

There are a lot of corpora available in English. Some of them merely capture a reader’s gaze behaviour as they read texts varying from articles to a novel. Others capture the reader’s gaze movement as the readers solve different NLP tasks. Along with English, there are a number of other publicly available corpora in other languages, like Chinese [Zang et al., 2018; Li et al., 2018], Dutch [Cop et al., 2017; Mak and Willems, 2019], German [Nicenboim et al., 2016; Kliegl et al., 2004] Persian [Safavi et al., 2016], Russian [Laurinavichyte et al., 2017], and Spanish [Nicenboim et al., 2016]. Table 1 gives the statistics for each of the publicly available eye-tracking corpora.

4 NLP Tasks Where Gaze Behaviour is Used

Gaze behaviour is also beneficial for multiple tasks in NLP where the reader’s psychological input is critical, and text features alone will not be enough. Mishra et al. [2013] discuss how using gaze behaviour would be a better approach to judge the complexity of translating a sentence, compared to using just length-based statistics (like word length, sentence length, etc.). Understanding sarcastic texts could also be resolved using gaze behaviour, where incongruity in the text (one of the leading indicators of sarcasm) induces gaze behaviour characterized by longer fixations, regressions etc. [Mishra et al., 2016a]. Gaze behaviour has also been used to identify a reader’s native language [Berzak et al., 2017], as well as in detecting grammatical errors in compressed sentences [Klerke et al., 2015b]. Klerke et al. [2015a] also show that gaze behaviour can be used to evaluate the output of Machine Translation systems better than automated metrics. Gaze behaviour has also been used to evaluate how a reader would rate the quality of a piece of text [Mathias et al., 2018].

The scanpath of a reader (i.e. the path a reader’s eye traverses when they read the text) has been used to test how easy/difficult a piece of text is for a reader to read [Mishra et al., 2017]. It can also be used to predict the misreadings among children with reading difficulties [Bingel et al., 2018].

Gaze behaviour has also been used in multiple areas of sentiment analysis such as sarcasm detection [Mishra et al., 2016b], sarcasm understanding [Mishra et al., 2016a], and sentiment analysis annotation tasks [ Joshi et al., 2014]. Klerke and Plank [2019] tackle the classical problem of POS (Part-of-speech) tagging with gaze data and provide a systematic overview of the influence of two independent levels of gaze data aggregation on low-level syntactic labelling tasks at two separate levels of complexity; i.e., a simple chunk boundary tagging and a supervised POS tagging task.

5 Learning Gaze Behaviour

As mentioned in the previous section, using gaze behaviour helps systems in solving many NLP tasks. However, using gaze behaviour, while alleviating the need to record it at test / run time, is still a relatively new challenge. The biggest benefit of learning gaze behaviour at run time is that we alleviate the requirement of reading the texts in order to get gaze behaviour information. In this section, we look at different tasks and different systems which alleviate the need for collecting gaze behaviour at run time.

5.1 NLP Tasks

Predicting Fixations while Reading Nilsson and Nivre [2009] describe an approach to detect which tokens readers fixate on while reading, using a transition-based approach. They use a transition-based model of reader’s eye movements to predict the next word that is fixated. They used features like token length, token frequency class, next token length, next token frequency class, etc. Matthies and Søgaard [2013] improve on their approach using a linear CRF model to determine which words are fixated on while reading. The features that they use are word length (for a window of 5 words), and word probability (for a window of 3 words). While fixation accuracy and F1 are comparable to Nilsson and Nivre [2009]’s approach, Matthies and Søgaard [2013]’s approach predicts the eye movements better for new readers, than when training and test data come from the same reader.

Predicting Grammatical Functions Barrett and Søgaard [2015b] describe an approach to predict the grammatical functions of words in a sentence (i.e. whether they are subject, object, etc.). They trained their system using gaze behaviour data from Barrett and Søgaard [2015a], and found statistically significant improvements over their baseline, using a L2-regularized logistic regression classifier.

Text Simplification Klerke et al. [2016] describe a way to simplify text by compressing sentences using gaze behaviour learnt at run time. They used the Ziff-Davis [Knight and Marcu, 2002], Broadcast [Clarke and Lapata, 2006], and Google [Filippova et al., 2015] datasets and multi-task learning, where the primary task is compressing the sentences, and the auxiliary tasks are learning the gaze behaviour attributes, namely the first fixation duration, and regression duration. They describe two multi-task learning models for using gaze behaviour. The first approach (called Multi-task in their paper) uses multi-task learning with a separate logistic
regression classifier for learning gaze behaviour during training. The second approach (Cascaded) also uses multi-task learning with gaze behaviour learnt from an inner layer in their network.

**Part-of-Speech (PoS) tagging** Barrett *et al.* [2016a] describe an approach to solve the task of part-of-speech tagging using gaze behaviour from the Dundee Treebank [Barrett *et al.*, 2015]. Barrett *et al.* [2016b] investigate the same using cross-lingual approaches (i.e., training on English, test on French, and vice versa). Both approaches use a Hidden Markov Model with additional type-aggregated gaze features. The features that they used were early gaze features (which measure events when the reader first encounters a word, like first fixation duration), late gaze features (which measure events when the reader finishes reading the word, like number of regressions to a word), basic gaze features (measures which don’t specifically belong to either the early or late gaze features, like dwell time). Regression-from features (which capture regressions from a word, like total regression from duration), context gaze features (namely the fixation durations and probabilities of neighbouring words), and finally, the NoGaze features, which are text features from the British National Corpus and Dundee Corpus [Kennedy *et al.*, 2003].

**Readability** González-Garduno and Søgaard [2018] describe a solution for predicting readability using multi-task learning, with readability prediction as the primary task, and predicting gaze behaviour as an auxiliary task. They use a multi-task multi-layer perceptron and multi-task logistic regression systems, with their best results achieved by using all features with the multi-task multi-layer perceptron, with their gaze behaviour learnt from the Dundee Corpus [Kennedy *et al.*, 2003].

**Sentiment Analysis** Mishra *et al.* [2018] describe a way to use gaze behaviour for the task of sentiment analysis of movie reviews. They use a multi-task framework, where the primary task is predicting the sentiment, and the auxiliary tasks are part-of-speech tagging and learning gaze behaviour. They use a pair of bi-directional LSTM layers, where each layer performs a particular task (either PoS tagging or learning gaze behaviour). They showed statistically significant improvements over the state-of-the-art reported results using both their multi-task systems on the IMDB25K dataset [Maas *et al.*, 2011], and the PL2000 dataset [Pang and Lee, 2004].

**Named Entity Recognition (NER)** Hollenstein and Zhang [2019] describe an approach to perform named entity recognition, using gaze behaviour aggregated over their gaze corpora showing improvements on the CoNLL 2003 dataset [Tjong Kim Sang and De Meulder, 2003]. To train their system to learn the gaze behaviour, they used the Dundee Corpus [Kennedy *et al.*, 2003], GECO Corpus [Cop *et al.*, 2017], and ZuCo Corpus [Hollenstein *et al.*, 2018] for collecting the gaze data. For each token, they extract gaze features from the gaze corpora using type aggregation. Similar to Barrett *et al.* [2016a], Hollenstein and Zhang [2019] use similar gaze features - namely early, late, basic and context gaze features. For each word, their system takes character and word embeddings, along with the gaze features as input, and returns the corresponding NER tag. To do this, they use a bi-directional

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
<th>Language</th>
<th>Stimulus</th>
<th>Total Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zang et al. (2018)</td>
<td>[Zang et al., 2018]</td>
<td>Chinese</td>
<td>90 sentences, 15 documents</td>
<td>35</td>
</tr>
<tr>
<td>Li et al. (2018)</td>
<td>[Li et al., 2018]</td>
<td>Chinese</td>
<td></td>
<td>29</td>
</tr>
<tr>
<td>GECO</td>
<td>[Cop et al., 2017]</td>
<td>Dutch</td>
<td>1 novel, 3 short stories</td>
<td>33</td>
</tr>
<tr>
<td>Mak &amp; Willems (2019)</td>
<td>[Mak and Willems, 2019]</td>
<td>Dutch</td>
<td></td>
<td>102</td>
</tr>
<tr>
<td>ZuCo</td>
<td>[Hollenstein et al., 2018]</td>
<td>English</td>
<td>1107 sentences, 1 novel, 40 documents</td>
<td>12</td>
</tr>
<tr>
<td>GECO</td>
<td>[Cop et al., 2017]</td>
<td>English</td>
<td>55 texts, 27 documents</td>
<td>84</td>
</tr>
<tr>
<td>Parker et. al. (2017)</td>
<td>[Parker et al., 2017]</td>
<td>English</td>
<td>30 documents, 32 documents</td>
<td>20</td>
</tr>
<tr>
<td>ASD Data</td>
<td>[Yaneva, 2016]</td>
<td>English</td>
<td>1000 tweets, 7</td>
<td>7</td>
</tr>
<tr>
<td>CFILT-Quality</td>
<td>[Mathias et al., 2018]</td>
<td>English</td>
<td>1059 sentences, 5</td>
<td>5</td>
</tr>
<tr>
<td>CFILT-Scanpath</td>
<td>[Mishra et al., 2017]</td>
<td>English</td>
<td>205 sentences, 43</td>
<td>43</td>
</tr>
<tr>
<td>CFILT-Coreference</td>
<td>[Cheri et al., 2016]</td>
<td>English</td>
<td>20 documents, 10</td>
<td>10</td>
</tr>
<tr>
<td>CFILT-Sarcasm</td>
<td>[Mishra et al., 2016a]</td>
<td>English</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CFILT-Sentiment</td>
<td>[Joshi et al., 2014]</td>
<td>English</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCL Corpus</td>
<td>[Frank et al., 2013]</td>
<td>English</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dundee</td>
<td>[Kennedy et al., 2003]</td>
<td>English</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dundee</td>
<td>[Kennedy et al., 2003]</td>
<td>French</td>
<td>20 documents, 10</td>
<td>10</td>
</tr>
<tr>
<td>Self-Paced Reading Time</td>
<td>[Nicenboim et al., 2016]</td>
<td>German</td>
<td>176 sentences, 72</td>
<td>72</td>
</tr>
<tr>
<td>Potsdam Sentence Corpus</td>
<td>[Kliegl et al., 2004]</td>
<td>German</td>
<td>144 sentences, 55</td>
<td>55</td>
</tr>
<tr>
<td>Dependency Resolution dataset</td>
<td>[Safavi et al., 2016]</td>
<td>Persian</td>
<td>136 sentences, 40</td>
<td>40</td>
</tr>
<tr>
<td>Russian Sentence Corpus</td>
<td>[Laurinavichyute et al., 2017]</td>
<td>Russian</td>
<td>144 sentences, 96</td>
<td>96</td>
</tr>
<tr>
<td>Self-Paced Reading Time</td>
<td>[Nicenboim et al., 2016]</td>
<td>Spanish</td>
<td>212 sentences, 79</td>
<td>79</td>
</tr>
</tbody>
</table>

Table 1: Summary of eye-tracking datasets available today. The novel read in the GECO dataset is *The Mysterious Affair at Styles* by Agatha Christie (in both Dutch and English). Dataset is the name of the dataset. Source is the reference where the dataset was published. Language is the language of the data in the dataset. Stimulus is the amount of text data that the annotators read. Total subjects is the number of subjects who participated in the creation of the corpus.
LSTM neural network with conditional random fields to perform NER.

**Sequence Classification**  Barrett et al. [2018] describe a solution to multiple NLP tasks - sentiment analysis, grammatical error detection, and hate speech detection - using multi-task learning. They use a bi-directional LSTM [Hochreiter and Schmidhuber, 1997] layer with attention for encoding each token to get a sentence representation. The network’s attention weights are learnt using existing gaze behaviour data, so that no gaze data recording is done at run time. They show that biasing the network’s attention using gaze behaviour leads to performance improvements across all 3 tasks as well as aid in explaining the network’s decisions.

**5.2 System Architectures**

One of the earliest frameworks to predict gaze behaviour was EZ-Reader [Reichle et al., 2003], which was used for understanding how word identification, visual processing, attention, and oculomotor control jointly determine when and where the eyes move during reading. Another system, called Swift [Engbert et al., 2005] used a mathematical model for the control of eye movements during reading that is both psychologically and neurophysiologically plausible and that accounts for most of the then known experimental findings.

There are other different approaches used today to alleviate the need for recording gaze behaviour at run time. Feature engineering is a way in which we predict the gaze behaviour data using lexical properties, such as word length, syllable count, word surprisal, etc. Type aggregation is another way in which we avoid tracking a reader’s eyes when they read the text. In this approach, we use a gaze corpus (like the Dundee Corpus) and pre-train our model with aggregate values of each gaze feature. Multi-task learning is an approach where learning gaze behaviour is an auxiliary task. This approach is often used to learn gaze at the word-level, but our task requires an output at the sentence/document level.

**5.3 Normalizing Data**

While collecting gaze behaviour data, it is better to normalize gaze behaviour across readers. This is done because readers may read the text either faster, or slower, depending on their fluency with the language. Therefore, we must normalize the gaze behaviour data before use. Gaze behaviour is normalized in the following ways. **Min-Max Normalization** is where we normalize the data on a scale of \([0, 1]\), where 0 corresponds to the lowest value, and 1 corresponds to the highest value. [Barrett et al., 2016a] is one work that uses min-max normalization. **Binning** is where we consider discrete bins for each gaze behaviour attribute. For example: 6 bins (numbered 0 to 5) as described by [Klerke et al., 2016]. One of the advantages of binning over min-max normalization is that it reduces the effect of outliers in the data. In both approaches of normalization, the normalization takes place for each reader.

**6 Applications of Learning Gaze Behaviour**

In this section, we cover some applications in NLP which we believe could benefit a lot from learning gaze behaviour. To the best of our knowledge, there is no work on any of these applications where the system learns gaze behaviour.

**6.1 Complex Word Identification**

Lexical simplification is a process in which text gets simplified by replacing complex words and phrases with simpler ones. For example, a non-native speaker of English will struggle to understand what the word “procrastinate” means, in the sentence “We should not procrastinate our submission.” However, they will be more likely to understand the meaning of the sentence “We should not delay our submission.” The process of identifying which words are hard (and should be replaced by an appropriate synonym) is a very useful application for the use of eye-tracking, as readers fixate longer on harder words compared to easier words [Rayner, 1998].

Paetzold and Specia [2016] report the results of the shared task on complex word identification held at Sem-Eval 2016. Another shared task was organized in 2018, to identify complex words, and phrases, in English, Spanish, German, and a cross-lingual setting (where the target language was French) [Yimam et al., 2018].

Kunze et al. [2013] showed how eye tracking can be used for identifying complex words. However, their approach required readers to read the text, while they track their eye movements, in order to identify complex words. Therefore, using cognitive information should aid in solving these tasks. As mentioned earlier, there are quite a few papers that deal with text complexity/simplification and readability which use gaze behaviour like Klerke et al. [2016], Mishra et al. [2017], González-Garduño and Søgaard [2017], González-Garduno and Søgaard [2018], etc. However, in proposing solutions for the shared tasks, none of them used cognitive information. With the availability of a large number of eye-tracking corpora, in various languages, an interesting avenue of research would be exploring how using gaze behaviour can help in identifying complex words, even if we don’t have such information at run time.

**6.2 Automatic Essay Grading**

An essay is a text, written in response to a topic, called the essay prompt. Grading an essay is assigning a score to the essay based on its quality, either for the essay on the whole (holistic scoring) or for certain aspects of the essay (trait-specific scoring) [Ke and Ng, 2019]. Automatic essay grading (AEG) is the process of grading an essay using a machine. The earliest AEG system was described by Page [1966], over 50 years ago. Since then, there have been a number of commercial AEG systems, like E-Rater [Attali and Burstein, 2006], Intelligent Essay Assessor [Landauer, 2003], LightSide [Mayfield and Rosé, 2013], etc.

The current state-of-the-art AEG systems use neural networks, like CNNs [Dong and Zhang, 2016], LSTMs [Alikaniotis et al., 2016; Taghipour and Ng, 2016; Tay et al., 2018], or both [Dong et al., 2017; Zhang and Litman, 2018; Liang et al., 2018]. The dataset that they use is the 2012 Automatic Student Assessment Prize (ASAP) AEG Dataset, released by the Hewlett Foundation [Ke and Ng, 2019].

As mentioned earlier, Mathias et al. [2018] describe a way to predict the rating a reader would give a piece of text based
on its quality. Their work showed that gaze behaviour we could use for similar applications like AEG. However, their approach required readers to read the text in order to use the reader’s gaze behaviour. An AEG system built using multi-task learning where the primary task is scoring the essay at the document-level, and learning the gaze behaviour is the auxiliary task at the word-level. Barrett et al. [2018] have shown this approach to benefit multiple NLP tasks like sentiment analysis, grammatical error detection and hate speech detection.

7 Conclusion

Gaze behaviour has been shown to aid multiple natural language processing tasks [Mishra and Bhattacharyya, 2018]. However, collecting gaze behaviour at run time is not feasible. Hence, in order to use gaze behaviour, we utilize different approaches, like multi-task learning, using type-aggregated values, etc.

In this paper, we first introduce the AI audience to different NLP tasks which are solved using gaze behaviour, like translation complexity, sarcasm understandability, text quality rating prediction, etc. We then discuss different tasks where we show that gaze behaviour aids in their solution. To solve any of these tasks, we require gaze behaviour data to be ready for training. In our paper, we also report gaze behaviour datasets created in multiple languages. Finally, we describe a pair of applications from the domain of education - complex word identification, and automatic essay grading - which could benefit a lot from using gaze behaviour based solutions.

References


