Sequence Labelling and Sequence Classification with Gaze: Novel Uses of Eye-Tracking Data for Natural Language Processing

Abstract

Eye-tracking data from reading provides a structured signal with a fine-grained temporal resolution which closely follows the sequential structure of the text. It is highly correlated with the cognitive load associated with different stages of human, cognitive text processing. While eye-tracking data has been extensively studied to understand human cognition, it has only recently been considered for Natural Language Processing (NLP). In this review, we provide a comprehensive overview of how gaze data is being used in data-driven NLP, in particular for sequence labelling and sequence classification tasks. We argue that eye-tracking may effectively counter one of the core challenges of machine-learning-based NLP: the scarcity of annotated data. We outline the recent advances in gaze-augmented NLP to discuss how the gaze signal from human readers can be leveraged while also considering the potentials and limitations of this data source.

Keywords: eye tracking, gaze, natural language processing, natural reading, human text processing, sequence labelling, sequence classification

Sequence Labelling and Sequence Classification with Gaze: Novel Uses of Eye-Tracking Data for Natural Language Processing

Introduction

During normal, skilled reading, the eyes move sequentially through the text, fixating one word at a time. In numerous, controlled psycholinguistic studies, word-based eye movement metrics have proven to be strongly correlated with high-level text processing, such as syntactic and semantic structures (Rayner, Sereno, Morris, Schmauder, & Clifton Jr, 1989).

Natural Language Processing (NLP) is an interdisciplinary field of linguistics and computer science that, e.g., tries to solve sequence labelling and sequence classification tasks. Such tasks are largely accessed automatically when humans read, e.g., named entity recognition and syntactic analysis. Recently, NLP has started to discover the potentials of gaze data for improving the performance of machine learning models. Such data is referred to as *fortuitous data* by Plank (2016b). This term covers "non-obvious data that is hitherto neglected, hidden in plain sight or raw data that needs to be refined" and is suggested to be leveraged when annotated resources are scarce.

This review is divided into four sections. We will first introduce basic concepts concerning eye-tracking data. Then, it provides a summary of the largest available gaze resources of naturalistic reading in English. The main part of the article is a comprehensive overview of recent advances in NLP concerning sequence labelling and sequence classification using gaze data. Finally, we will summarise observations across the studies of the survey on how to include eye movements in NLP and discuss the potentials and limitations of the data source.

Scope

The scope of this review is data-driven NLP, limited to sequence labelling and sequence classification tasks on English text using eye movements from adults performing naturalistic reading. Naturalistic reading denotes reading of naturally

occurring text without time constraints, task solving[1] or other reading constraints, such as limiting the preview of the following word. Therefore, we do not cover the fairly large body of work on annotators' eye movements (see review by Mishra and Bhattacharyya (2018)).

We will focus on studies where an actual *evaluation* of an NLP sequence labelling or sequence classification task has taken place. At the beginning of each section, we will present and discuss the experimental or large-scale psycholinguistic studies that *describe* correlations between eye movements and the linguistic phenomena under consideration. This review will not cover dialogue, studies modelling reader attributes from eye movements, or information retrieval studies.

A Very Brief Introduction to Eye Movements

Contrary to the perceived experience, the reader's eyes do not glide smoothly across the lines of the text while reading. Instead, the eye movements alternate between fixating regions of the text and performing rapid, ballistic movements, named saccades. The fixations last on average 200-250ms, with large temporal variations. Conversely, saccades are much shorter, lasting around 20-40ms and typically move the eyes 7-9 characters forward. During saccades, the eye does not take in information. In the course of normal reading, around 10% of the saccades, often unconscious to the reader, move back to a previously read part of the text for further processing, and possibly re-processing. These movements are called regressions. Eye movements thereby allow studying early and late cognitive processing separately using a range of well-established word-based metrics. Eye movements represent therefore one of the richest behavioural data source for human text processing during reading. The section "Other Human Data Sources" presents a comparison to other behavioural measures in NLP.

Moreover, it is useful to differentiate between low- and high-level factors both influencing how the eyes act during reading. Low-level processing accounts for how the eyes perceive and decode text. High-level processes encompass syntactic and semantic

[1] In some studies, subjects may answer comprehension questions or solve other tasks after reading, but when this is separated from the reading process, we still consider it naturalistic reading.

processing, which are of interest to NLP. However, the interaction between low and high-level factors should be included in NLP models.

Low-Level Processing

Due to the anatomy of the eye, readers can only see a small part of the text during each fixation. This asymmetric area extends, for a skilled reader of an alphabetic language arranged from left to right, 3-4 characters to the left of the fixation point and 14-15 characters to its right (Rayner, Well, & Pollatsek, 1980). In practice, words are identified in an even smaller area extending merely 7-8 characters to the right of the fixation point, called *the perceptual span* (McConkie & Rayner, 1976). The saccade distance and the size of the perceptual span vary as a function of text difficulty and reading skill.

High-Level processing

Fixation durations are shorter if a word is easy to identify and understand (Clifton, Staub, & Rayner, 2007). "Easy to identify and understand" extends over a large range of high-level effects, e.g., predictability from context (Inhoff, 1984), age of acquisition (Kemper & Liu, 2007), familiarity (Juhasz & Rayner, 2003; Williams & Morris, 2004), and morphology (Hyönä, Bertram, & Pollatsek, 2004). One of the most studied effects is the influence of word frequency on fixation durations. It has been observed that readers tend to look longer at infrequent words and/or long words (Rayner, 1977). Nearly 70% of the variance in mean fixation duration can be explained by word length and word frequency (Just & Carpenter, 1980).

What can be learned from the basic understanding of eye movements?

When not in a controlled experimental setup, one challenge about eye movements is that they reflect cognitive processing, but not *which process*. Hence, gaze is an *indirect* measure of cognitive processing. It will, therefore, help machine learning models to identify the signal in the eye movements if confounding high- and low-level effects are also included. Due to spill-over and preview effects, it is also useful to provide

information about the previous and subsequent words and fixations when using a model where this is not already accounted for.

Large, English Eye-Tracking Corpora

The text stimulus used in most psycholinguistic studies consists of constructed sentences that occur infrequently in natural text and that are often read out of context. This is not favourable for NLP purposes that generally favour large quantities of naturally occurring text.

An overview of the available, large (>2000 words), English corpora of naturalistic reading of naturally occurring text by native speakers is provided in Table 1. Some corpora, such as the Dundee Corpus (Kennedy, Hill, & Pynte, 2003) and the GECO corpus (Cop, Dirix, Drieghe, & Duyck, 2017) also contained data in another language or read by non-native subjects. These data are omitted from the table.

Using Gaze for Sequence Labelling and Sequence Classification

This section is thematically ordered according to NLP topics, where relevant work on sequence labelling and sequence classification tasks using gaze data has been accomplished. Each subsection will begin with a summary of psycholinguistic findings, relevant to the considered topic, both experimental and on larger quantities of naturalistic reading data.

Text Complexity

Text complexity covers a range of textual features on global, syntactic and word level. It applies to well-edited text as well as irregularities. Furthermore it interact with reading skill, thereby linking it to text comprehension. Text comprehension has been thoroughly studied using eye movements (Rayner, Chace, Slattery, & Ashby, 2006; Vasishth, von der Malsburg, & Engelmann, 2013). Processing difficulties have an impact on regressions, saccade distance, and fixation durations. The psycholinguistic studies on text comprehension have motivated the use in NLP of gaze for the evaluation of text complexity. For example, gaze has been employed for scoring the readability,

grammaticality or acceptability on its own or in the context of evaluating the output from automatic systems.

Diverse machine learning approaches have allowed leveraging the correlation of fixation duration-based metrics and readability. Mathias et al. (2018) improved text quality evaluation by using gaze features to represent words. Specifically, they predicted text quality attributes (organization, coherence, cohesion, and quality) from fixation and regression features, as well as from textual features using a single-layer feed-forward neural network. Their best model combined gaze features with textual features. While directly using gaze as word-based features is the most straight-forward input method, it requires also the most eye-tracking data. The study by Mathias et al. (2018), therefore, required a study-specific data collection. Singh, Mehta, Husain, and Rajakrishnan (2016) used instead predicted gaze metrics learned from human reading as features to predict the readability, hence alleviating the need to collect new data for each application. A system was first trained on the Dundee Corpus (Kennedy et al., 2003) and used to predict gaze features on the target text. The predicted gaze features were then used as features in a logistic regression classifier to predict sentence complexity. Multitask learning (Caruana, 1997) represents another method that alleviates the need for gaze data on the target text. González-Garduño and Søgaard (2017) showed that multitask sentence-level text readability classification performed better than a single-task setup. Their best model was a multi-layer perceptron where readability and averaged total fixation duration were predicted simultaneously.

Evaluating the Complexity of Automatically Generated Text. Human evaluation is commonly used as a reference when exploring how to best evaluate the output from automatic systems (Chaganty, Mussmann, & Liang, 2018; Resnik & Lin, 2010). This is, however, very expensive. As a proxy, the reader's eye movements have been used instead as a signal to evaluate the quality of texts generated by computer systems, e.g., machine translations and automatic summarisations. In this context, gaze can be considered as noisy, human annotations. Results suggest that the eye movements reveal the location of machine translation errors (Bremin et al., 2010; Doherty &

O'Brien, 2009) and that in some cases gaze metrics may even help distinguish the type of error (Stymne et al., 2012). Klerke, Castilho, Barrett, and Søgaard (2015) found that reading metrics were better proxies for the usability of translated text than the standard metric, i.e., bilingual evaluation understudy score (Papineni, Roukos, Ward, & Zhu, 2002).

Klerke, Alonso, and Søgaard (2015) used gaze to evaluate the output of an automatic sentence compression system; grammatical as well as ungrammatical. The original, uncompressed sentence and a human-made compressed version of the sentences were included as references. They found that the reading speed and regression features were the most informative gaze features for detecting ungrammaticality in the automatically compressed sentences. In a multitask learning setup, including a gaze feature as an auxiliary task, improved the performance of a sentence compression system (Klerke, Goldberg, & Søgaard, 2016).

Scanpath Metrics. The scanpath is the trajectory of the eyes through the text. The scanpath over a text sequence has also been summarised in quantitative metrics, mainly for its expressivity of text comprehension problems. Scasim is a sentence-level score that was used to detect irregularities in the scanpath during reading (von der Malsburg & Vasishth, 2011). Mishra, Kanojia, Nagar, Dey, and Bhattacharyya (2017) defined another scanpath metric, scanpath complexity, that correlated with different measures of lexical and syntactic complexity as well as standard readability metrics. Though the task is framed as modelling reading effort, it is relevant for text complexity evaluation as well. Wallot, O'Brien, Coey, and Kelty-Stephen (2015) also showed that the degree of power-law scaling in raw eye movements was predictive of text comprehension.

Part-of-Speech

There is not a lot of literature concerning how the cognitive processing of word classes is reflected in the eye movements. But Carpenter and Just (1983) noted that 38% of function words are fixated and 83% of content words are fixated. There is,

however, evidence that the probability of a word class given the preceding context is negatively correlated with the fixation duration/probability (Bauman, 2013; Demberg & Keller, 2008; Pynte & Kennedy, 2007). For French, the gaze time on a target word depended on the degree of semantic relatedness to two nouns/adjectives/verbs belonging to a prior part of the sentence and located at varying distances. Only verbs were found to have an effect for the longest distance which suggest that nouns and adjectives exert their influence at a local level, and verbs at a more distant level (Pynte, New, & Kennedy, 2009). None of the above-mentioned studies tried to distinguish or characterise the reading of a broad range of part-of-speech (POS) classes, but they all showed that the processing of word classes is dependent on the context.

Results from NLP suggest that eye movements can disambiguate some word classes when gaze features were used as features in a supervised POS tagger (Barrett & Søgaard, 2015a). The tagger was a Perceptron-based model with dropout. Gaze features were used as multi-dimensional, continuous representations of each word type similarly to word embeddings. Similar representations were also useful for POS tagging when employed in an unsupervised sequence induction algorithm, i.e., a hidden Markov model (Barrett, Bingel, Keller, & Søgaard, 2016). A similar architecture was employed in Barrett, Keller, and Søgaard (2016) where results suggest that such gaze correlations may transfer across related languages. Here, English gaze data was used to improve POS induction for French. Klerke and Plank (2019) also found that predicting a gaze feature as an auxiliary task may help POS tagging a multitask learning setup.

Syntax

Clifton et al. (2007) presented an exhaustive survey over higher-level effects in psycholinguistic studies of human reading. Here, all four studies on syntactic complexity in sentences without syntactic ambiguity found an effect of syntactic complexity on early gaze measures (Hyönä & Vainio, 2001; Rayner, Sereno, et al., 1989; Staub, Clifton Jr, & Frazier, 2006; Vainio, Hyönä, & Pajunen, 2003). This was also supported by results from a large-scale study: Demberg and Keller (2008) found that a measure of

syntactic complexity, namely integration cost (Gibson, 2000), was positively correlated with fixation duration for nouns. Hence, intervening grammatical structures between the head and the dependant slowed down the parsing of the sentence during reading.

In the field of NLP, Barrett and Søgaard (2015b) showed that token-based eye movement features could, to some extent, disambiguate four syntactic roles for nouns using a logistic regression classifier. They also demonstrated that word-type-based gaze features helped supervised dependency parsing better than pre-trained word embeddings. Barrett, González-Garduño, Frermann, and Søgaard (2018) used gaze features combined with prosodic features on word-type level in an unsupervised sequence labelling algorithm, a hidden Markov model, for syntactic chunking. This representation was also found to be better than pre-trained word embeddings. Strzyz, Vilares, and Gómez-Rodríguez (2019a) combined many of the elements from these studies in a multitask recurrent neural network for dependency parsing using hard-parameter sharing. Here, dependency parsing was, rather unconventionally, treated as a sequence labelling problem for a recurrent neural network to model. Following Strzyz, Vilares, and Gómez-Rodríguez (2019b), they predicted two main tasks; the index of the head and the relation between the head and the dependent. The auxiliary task was also a sequence labelling problem of predicting one or more of 12 discretised gaze features (up to four at a time). An auxiliary task for each feature was instantiated. They experimented both with joint and disjoint data for the main task and the auxiliary task and found small, but consistent improvements for both setups. The best feature(s) varied on the test and the development set but overall mean fixation duration and context features (fixation duration/probability on the word before or after the target word) were most helpful.

Pragmatics: Sarcasm Detection

Inferring pragmatics from eye movements may depend on internal attributes of the reader, such as social knowledge, mental state and attentiveness to a higher degree than, e.g., syntactic processing. We nevertheless include this line of work, assuming

that the main objective of sarcasm detection is to learn attributes of the text and not the reader. There are several, conflicting cognitive hypotheses and evidence concerning the processing of irony, some saying that irony is always processed twice, and hence slower (Grice, 1975) and some saying that only in specific cases, irony is processed slower (Filik, Leuthold, Wallington, & Page, 2014; Gibbs Jr, 1994; Gibbs, 1986; Ivanko & Pexman, 2003).

There is evidence that eye movements can help to predict whether a reader caught the sarcastic meaning of a sentence or not. Most approaches explore the type of sarcasm that is related to incongruity with the context. For example, Mishra, Kanojia, and Bhattacharyya (2016), Mishra, Kanojia, Nagar, Dey, and Bhattacharyya (2016a) used a set of scanpath-based features on sentence-level for binary sarcasm classification, and show improvements over non-gaze features by supervised machine learning algorithms. The best model used gaze features along with textual features. All approaches relied, however, on manual feature engineering. To overcome the problem of manual feature engineering, features from the scanpath were used to train a convolutional neural network for sarcasm classification (as well as sentiment classification)(Mishra, Dey, & Bhattacharyya, 2017). The learned features outperformed the manually engineered gaze features used by Joshi, Sharma, and Bhattacharyya (2015), Mishra et al. (2016a), and a baseline convolutional neural network which only relied on the textual input.

Named Entity Recognition and Relations in Text

There are various pieces of evidence in favour of using eye-tracking data for named entity recognition (NER): word familiarity and predictability had a negative effect on fixation duration. Additionally, reading patterns contain indications of syntactical categories (see the section “Part-of-Speech”). This indicates that the reading of unfamiliar proper nouns (such as names for persons, organisations and locations, i.e. named entities) may have a distinct reading pattern. Tokunaga, Nishikawa, and Iwakura (2017) analysed eye tracking signals during the annotation of named entities in order to extract useful features for NER. Their work shows that humans took a broad

context into account to identify named entities, including predicate-argument structure. This hints to the usefulness of eye tracking recordings of full sentences for this task.

Hollenstein and Zhang (2019) leveraged eye movement data from three gaze corpora and concatenated character, word and gaze feature vectors as input to a recurrent neural network for named entity recognition.

Gaze may also be used to detect relations in text: results from the statistical analysis of Jaffe, Shain, and Schuler (2018) suggested that gaze might assist co-reference resolution because entities with more mentions are processed faster. Moreover Yaneva, Evans, Mitkov, et al. (2018) showed that eye movements were useful for classifying referential and non-referential uses of *it* and Cheri, Mishra, and Bhattacharyya (2016) showed that regression features specifically could improve coreference classifiers.

Multiword Expressions

Multiword expressions vary in their linguistic properties but they are perceived as highly conventional by native speakers (Siyanova-Chanturia, 2013). They pose challenges for, e.g., machine translation and it is, therefore, useful to detect them automatically. Multiword expressions are an example of eye movement processing on the super-word level. In behavioural eye-tracking experiments, the entire multiword expression was found to have a processing advantage over novel strings of language (Schmitt & Underwood, 2004; Yaneva, Taslimipoor, Rohanian, et al., 2017). Rohanian, Taslimipoor, Yaneva, and Ha (2017) found that multiword expressions could be predicted by a conditional random field from gaze features from both native and second language speakers equally, where late processing measures achieved better results than early measures.

Sentiment Classification and Other Sequence Classification Tasks

Detecting semantic characteristics of sentences and contextual connotations of words from eye movements is dependent on the subjectivity and general knowledge of a reader. Nevertheless, not only could eye- tracking features be used to improve sentiment analysis on the sentence level (Mishra, Kanojia, Nagar, Dey, & Bhattacharyya, 2016b),

but eye-tracking features could be learned directly from scanpaths (Mishra, Dey, & Bhattacharyya, 2017)[2]. Barrett, Bingel, Hollenstein, Rei, and Søgaard (2018), Long, Lu, Xiang, Li, and Huang (2017), Long, Xiang, Lu, Huang, and Li (2019) used gaze to weigh which words (and for the two former: also sentences) got more attention by a recurrent neural network classifier for sentiment classification. Barrett, Bingel, Hollenstein, et al. (2018) also showed that this approach worked for grammatical error detection and hate speech detection.

How to Use Gaze for NLP

This section contains an examination on how to use gaze for NLP based on observations across the NLP studies from the survey above.

Which Gaze Features?

All NLP studies that represent gaze as word representations use several gaze features. In this exhaustive survey, the reviewed studies which directly incorporate gaze as multidimensional, continuous features, use between 4 and 34 features.

Several NLP studies tried to identify the best combination of gaze features by systematically grouping them. In most studies, the best results were obtained by using all the gaze features, e.g., for multiword expression prediction (Rohanian et al., 2017), named entity recognition (Hollenstein & Zhang, 2019) and POS induction (Barrett, Bingel, Keller, & Søgaard, 2016). When studying the contribution of individual gaze features, Barrett and Søgaard (2015a, 2015b), Mishra et al. (2016b) found that the signal is distributed over many eye movement features for classification of POS, grammatical function, sentiment, and sarcasm. The optimal eye-tracking features depend on the task but for complex phenomena, the cognitive processing seems to be distributed over a wide range of word-level, eye-tracking features.

In all studies that did not use deep neural architectures, the eye movement features were always combined with textual features. The textual features appear to supplement eye movements, thus improving the performance of the model. Mishra et al.

[2]This approach is already introduced in the section “Sarcasm”

(2016b) systematically combined eye movement features with sentiment, sarcasm, irony, and thwarting related features, and features related to reading difficulty. They find that combining all feature groups overall improve the prediction of sarcasm and sentiment. Similarly, Yaneva, Evans, Mitkov, et al. (2018) increased the accuracy for classifying referential uses of *it* by combining eye movements combined with linguistic features. The best model in Barrett, González-Garduño, et al. (2018) for POS induction used pre-trained word embeddings combined with eye movement features.

Deep recurrent neural architectures automatically learn the textual feature weights, so for these models, manually engineered textual features are not required (e.g., Hollenstein and Zhang (2019), Strzyz et al. (2019a)).

Mishra, Dey, and Bhattacharyya (2017) presented a promising approach where the gaze representation is learned in a convolutional neural network from the raw scanpath instead of – as all the remaining studies in the survey – relying on manually selected gaze features. This approach yielded better performance for sentiment and sarcasm detection than using manually selected features.

How to Include Gaze Features for Training and Testing

There are several ways to include eye movements in NLP models, some of which also alleviate the need to have gaze features at test time. Currently, limited amounts of available eye-tracking data restrict the training and evaluation of NLP models.

The gaze features can simply be added to the word-based features as multi-dimensional vectors representing each word (Barrett & Søgaard, 2015a; Rohanian et al., 2017; Yaneva, Evans, Mitkov, et al., 2018). However, this approach would also require gaze features for the test set which does not scale well to real-world applications. Barrett, Bingel, Keller, and Søgaard (2016), Barrett, Keller, and Søgaard (2016) showed that word-type averages of gaze features helped POS induction better than token-level features. Klerke and Plank (2019) found that word-type variance was better than less aggregated gaze features. Using word-type gaze features does not require gaze at test time. The features can be used similarly to word embeddings and

several studies also successfully concatenated type-level gaze features with pretrained word embeddings for a richer representation (Barrett, González-Garduño, et al., 2018; Hollenstein & Zhang, 2019; Mathias et al., 2018).

There are reliable rule-based reader models predicting eye movement behaviour, e.g., Reichle, Pollatsek, Fisher, and Rayner (1998). But the following studies each trained a machine-learning-based reader model to get predicted gaze-annotation for the task data without performing a new data collection (Long, Lu, et al., 2017; Long, Xiang, et al., 2019; Singh et al., 2016).

Multitask learning combines what is learned about the main task with what is learned from the gaze signal using parameter sharing between the tasks. Predicting gaze as an auxiliary task in a multitask learning setup is yet another approach that leverages the gaze signal without needing gaze-annotation of the main task data (González-Garduño & Søgaard, 2017; Klerke, Goldberg, & Søgaard, 2016; Klerke & Plank, 2019; Strzyz et al., 2019a). These studies employ a multitask learning setup for text compression, readability prediction, syntactic tagging, and dependency parsing respectively, while also learning to predict one or more gaze features. Only Strzyz et al. (2019a) employed several auxiliary tasks in the same model. Their best model predicted 4 gaze features: fixation duration and fixation probability of the previous and next word.

A similar approach consists in regularising the attention with gaze during the training of a recurrent neural network. Not all words are equally important for sequence classification and gaze durations may give cues about which words are more important. Long, Lu, et al. (2017), Long, Xiang, et al. (2019) directly distributed the attention from the (predicted) gaze durations. Barrett, Bingel, Hollenstein, et al. (2018) is in practice related to multitask learning since they performed alternate training. The main task batches were used to update the main parameters of the model and the gaze batches were used to update the attention weights of a recurrent neural network. Therefore, the gaze-annotated data and the task data were disjoint but the two tasks were trained in parallel, much like multitask learning.

Potentials for gaze in NLP

Even though many studies in this survey use eye movements in supervised models on more or less canonical text, we believe that the biggest potential for this data source is elsewhere. We attribute the use of canonical text to the scarcity of large gaze-annotated resources and we credit the use of supervised models to the fact that companionship between eye movements and NLP is fairly new. We agree with Plank (2016b) that there is unused potential in fortuitous data, such as gaze data, for non-canonical language as well as for low-resource languages.

There are more than 7000 languages in the world[3], and only a few of them have annotated resources to train supervised models (Plank, 2016b). It is faster and cheaper to have skilled native speakers read a text than professional annotators to annotate it. Furthermore, trained annotators may be hard to find for some low-resource languages. Eye trackers are increasingly available at a lower cost which seems promising for the availability of larger quantities of eye-tracking data (Krafka et al., 2016; San Agustin et al., 2010). This survey contains evidence that eye movements from skilled readers contain traces of human cognitive processing of linguistic phenomena that NLP models struggle to learn. The signal can also be leveraged by unsupervised algorithms; alone (Barrett, Bingel, Keller, & Søgaard, 2016; Barrett, Keller, & Søgaard, 2016) or combined with word embeddings or other accessible human text processing features (Barrett, González-Garduño, et al., 2018).

Other Human Data Sources

Other human data sources capturing cognitive processing have proven useful for improving NLP. This section will provide a brief overview of the attempts to improve NLP with other human data sources and compare them to eye movements.

Self-paced reading times are shallow and cheaper alternatives to eye movements. Enochson and Culbertson (2015) found that crowd-sourced reading times were comparable in quality to reading times recorded in a laboratory making this an

[3]https://www.ethnologue.com/

affordable and promising data source. Decision times, keystroke metrics, and speech can also – opposed to eye-tracking data – be collected with current consumer technology and have also shown useful for NLP, though not across as many tasks as NLP. Also, their link to the cognitive processing of text is less documented than it is for eye-tracking data. For instance, Plank (2016a) used keystroke logs to aid parsing and shallow parsing and Barrett, González-Garduño, et al. (2018), Pate and Goldwater (2011, 2013) used acoustic cues/prosodic features for parsing/syntactic chunking.

Direct measures of brain activity have been employed to improve NLP, especially neuroimaging techniques such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI). There are still very few EEG datasets which are usable for NLP. The ZuCo corpus (Hollenstein, Rotsztejn, et al., 2018) is an exception and provides simultaneous eye-tracking and EEG recordings. fMRI data has been used more widely than EEG: Wehbe, Vaswani, Knight, and Mitchell (2014) recorded data from subjects reading stories and aligned statistical language models with brain activity and Bingel, Barrett, and Søgaard (2016) extracted token-level signals of syntactic processing from fMRI for POS induction. Although the spatial accuracy of neuroimaging data sources might be beneficial, they are noisier, more expensive and more cumbersome to acquire. Moreover, while eye-tracking and EEG data easily allow for word-level data recordings in natural reading, this is less trivial for the low temporal resolution of other neuroimaging techniques, such as fMRI. Hence, the potentials of eye-tracking data, in addition to their accurate representation of many linguistic features, also lie in their relative effortless recording method and their higher signal-to-noise ratio.

Conclusion

Many of the tasks that NLP systems try to learn are largely accessed automatically when humans read. The cognitive processing of text has been studied carefully via eye movements. We presented an overview of recent NLP applications for sequence labelling and sequence classification that harness human cognition through the

use of eye movement features. The usages span over a range of NLP tasks: semantic, syntactic, POS, and relational.

The expected increase of availability of eye-tracking data should further encourage the usage of gaze data in NLP and will allow for researching additional methods on how to include it in machine learning algorithms. The biggest prospects are in non-canonical language and low-resource languages. One of the potentials includes investigating further how the data bottleneck of machine learning can be alleviated.

References

Barrett, M., Agić, Ž., & Søgaard, A. (2015). The Dundee Treebank. In *The 14th international workshop on treebanks and linguistic theories*.

Barrett, M., Bingel, J., Hollenstein, N., Rei, M., & Søgaard, A. (2018). Sequence classification with human attention. In *Proceedings of the 22nd conference on computational natural language learning* (pp. 302–312).

Barrett, M., Bingel, J., Keller, F., & Søgaard, A. (2016). Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th annual meeting of the association for computational linguistics* (Vol. 2, pp. 579–584).

Barrett, M., González-Garduño, A. V., Frermann, L., & Søgaard, A. (2018). Unsupervised induction of linguistic categories with records of reading, speaking, and writing. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies* (Vol. 1, pp. 2028–2038).

Barrett, M., Keller, F., & Søgaard, A. (2016). Cross-lingual transfer of correlations between parts of speech and gaze features. In *Proceedings of the 26th international conference on computational linguistics: technical papers* (pp. 1330–1339).

Barrett, M. & Søgaard, A. (2015a). Reading behavior predicts syntactic categories. In *Proceedings of the nineteenth conference on computational natural language learning* (pp. 345–249).

Barrett, M. & Søgaard, A. (2015b). Using reading behavior to predict grammatical functions. In *Proceedings of the sixth workshop on cognitive aspects of computational language learning* (pp. 1–5).

Bauman, P. (2013). Syntactic category disambiguation within an architecture of human language processing. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35, *35*, pp. 1833–1838).

Bingel, J., Barrett, M., & Søgaard, A. (2016). Extracting token-level signals of syntactic processing from fMRI - with an application to PoS induction. In *Proceedings of*

the 54th annual meeting of the association for computational linguistics (Vol. 1, pp. 747–755).

Bremin, S., Hu, H., Karlsson, J., Lillkull, A. P., Wester, M., Danielsson, H., & Stymne, S. (2010). Methods for human evaluation of machine translation. *Small*, *14*, 55–67.

Carpenter, P. A. & Just, M. A. [Marcel Adam]. (1983). What your eyes do while your mind is reading. *Eye movements in reading: Perceptual and language processes*, 275–307.

Caruana, R. (1997). Multitask learning. *Machine learning*, *28*(1), 41–75.

Chaganty, A., Mussmann, S., & Liang, P. (2018). The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (Vol. 1, pp. 643–653).

Cheri, J., Mishra, A., & Bhattacharyya, P. (2016). Leveraging annotators' gaze behaviour for coreference resolution. In *Proceedings of the 7th workshop on cognitive aspects of computational language learning* (pp. 22–26).

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In *Eye movements: a window on mind and brain* (pp. 341–371). Amsterdam, The Netherlands: Elsevier.

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: an eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, *49*(2), 602–615.

Demberg, V. & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*, 193–210.

Doherty, S. & O'Brien, S. (2009). Can MT output be evaluated through eye tracking. *Proceedings of MT Summit XII*, 214–221.

Enochson, K. & Culbertson, J. (2015). Collecting psycholinguistic response time data using amazon mechanical turk. *plosONE*.

Filik, R., Leuthold, H., Wallington, K., & Page, J. (2014). Testing theories of irony processing using eye-tracking and erps. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(3), 811.

Gibbs Jr, R. W. (1994). Figurative thought and figurative language. In *Handbook of psycholinguistics* (pp. 411–446). New York, NY: Academic Press.

Gibbs, R. W. (1986). On the psycholinguistics of sarcasm. *Journal of Experimental Psychology: General*, *115*(1), 3.

Gibson, E. (2000). The dependency locality theory: a distance-based theory of linguistic complexity. *Image, language, brain*, 95–126.

González-Garduño, A. V. & Søgaard, A. (2017). Using gaze to predict text readability. In *Proceedings of the 12th workshop on innovative use of nlp for building educational applications* (pp. 438–443).

Grice. (1975). Logic and conversation. In *Speech acts: syntax and semantics* (pp. 41–58). New York, NY: Academic Press.

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., & Langer, N. (2018). Zuco, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific data*, *5*, 180291.

Hollenstein, N. & Zhang, C. (2019). Entity recognition at first sight: improving NER with eye movement information. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 1–10).

Hyönä, J., Bertram, R., & Pollatsek, A. (2004). Are long compound words identified serially via their constituents? evidence from an eyemovement-contingent display change study. *Memory & Cognition*, *32*(4), 523–532.

Hyönä, J. & Vainio, S. (2001). Reading morphologically complex clause structures in finnish. *European Journal of Cognitive Psychology*, *13*(4), 451–474.

Inhoff, A. W. (1984). Two stages of word processing during eye fixations in the reading of prose. *Journal of verbal learning and verbal behavior*, *23*(5), 612–624.

Ivanko, S. L. & Pexman, P. M. (2003). Context incongruity and irony processing. *Discourse Processes*, *35*(3), 241–279.

Jaffe, E., Shain, C., & Schuler, W. (2018). Coreference and focus in reading times. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics* (pp. 1–9).

Joshi, A., Mishra, A., Senthamilselvan, N., & Bhattacharyya, P. (2014). Measuring sentiment annotation complexity of text. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (Vol. 2, pp. 36–41).

Joshi, A., Sharma, V., & Bhattacharyya, P. (2015). Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing* (Vol. 2, pp. 757–762).

Juhasz, B. J. & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1312.

Just, M. A. [Marcel A] & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review*, *87*(4), 329–354.

Kemper, S. & Liu, C.-J. (2007). Eye movements of young and older adults during reading. *Psychology and Aging*, *22*(1), 84.

Kennedy, A., Hill, R., & Pynte, J. (2003). The Dundee Corpus. In *Proceedings of the 12th European conference on eye movement*.

Klerke, S., Alonso, H. M., & Søgaard, A. (2015). Looking hard: eye tracking for detecting grammaticality of automatically compressed sentences. In *Proceedings of the 20th Nordic conference of computational linguistics* (pp. 97–105).

Klerke, S., Castilho, S., Barrett, M., & Søgaard, A. (2015). Reading metrics for estimating task efficiency with MT output. In *Proceedings of the sixth workshop on cognitive aspects of computational language learning* (pp. 6–13).

Klerke, S., Goldberg, Y., & Søgaard, A. (2016). Improving sentence compression by learning to predict gaze. In *Proceedings of 14th annual conference of the North American chapter of the association for computational linguistics* (pp. 1528–1533).

Klerke, S. & Plank, B. (2019). At a glance: the impact of gaze aggregation views on syntactic tagging. In *Lantern@emnlp*.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., & Torralba, A. (2016). Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2176–2184).

Long, Y., Lu, Q., Xiang, R., Li, M., & Huang, C.-R. (2017). A cognition based attention model for sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 462–471).

Long, Y., Xiang, R., Lu, Q., Huang, C.-R., & Li, M. (2019). Improving attention model based on cognition grounded data for sentiment analysis. *IEEE Transactions on Affective Computing*.

Luke, S. G. & Christianson, K. (2018). The Provo corpus: a large eye-tracking corpus with predictability norms. *Behavior research methods*, 1–8.

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

Mathias, S., Kanojia, D., Patel, K., Agrawal, S., Mishra, A., & Bhattacharyya, P. (2018). Eyes are the windows to the soul: predicting the rating of text quality using gaze behaviour. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (Vol. 1, pp. 2352–2362).

McConkie, G. W. & Rayner, K. (1976). Asymmetry of the perceptual span in reading. *Bulletin of the psychonomic society*, *8*(5), 365–368.

Mishra, A. & Bhattacharyya, P. (2018). *Cognitively inspired natural language processing: an investigation based on eye-tracking*. Springer.

Mishra, A., Dey, K., & Bhattacharyya, P. (2017). Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network.

In *Proceedings of the 55th annual meeting of the association for computational linguistics* (Vol. 1, pp. 377–387).

Mishra, A., Kanojia, D., & Bhattacharyya, P. (2016). Predicting readers' sarcasm understandability by modeling gaze behavior. In *Thirtieth AAAI conference on artificial intelligence* (pp. 3747–3753).

Mishra, A., Kanojia, D., Nagar, S., Dey, K., & Bhattacharyya, P. (2016a). Harnessing cognitive features for sarcasm detection. In *Proceedings of the 54th annual meeting of the association for computational linguistics* (Vol. 1, pp. 1095–1104).

Mishra, A., Kanojia, D., Nagar, S., Dey, K., & Bhattacharyya, P. (2016b). Leveraging cognitive features for sentiment analysis. In *Proceedings of the 20th signll conference on computational natural language learning* (pp. 156–166).

Mishra, A., Kanojia, D., Nagar, S., Dey, K., & Bhattacharyya, P. (2017). Scanpath complexity: modeling reading effort using gaze information. In *Thirty-first AAAI conference on artificial intelligence*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318). Association for Computational Linguistics.

Pate, J. K. & Goldwater, S. (2011). Unsupervised syntactic chunking with acoustic cues: computational models for prosodic bootstrapping. In *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics* (pp. 20–29). Association for Computational Linguistics.

Pate, J. K. & Goldwater, S. (2013). Unsupervised dependency parsing with acoustic cues. *Transactions of the Association for Computational Linguistics*, *1*, 63–74.

Plank, B. (2016a). Keystroke dynamics as signal for shallow syntactic parsing. In *Proceedings of the 25th international conference on computational linguistics* (pp. 609–618).

Plank, B. (2016b). What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of the conference on natural language processing (KONVENS)* (pp. 13–20).

Pynte, J. & Kennedy, A. (2007). The influence of punctuation and word class on distributed processing in normal reading. *Vision Research*, *47*(9), 1215–1227.

Pynte, J., New, B., & Kennedy, A. (2009). On-line contextual influences during reading normal text: the role of nouns, verbs and adjectives. *Vision research*, *49*(5), 544–552.

Rayner, K. (1977). Visual attention in reading: eye movements reflect cognitive processes. *Memory & Cognition*, *5*(4), 443–448.

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific studies of reading*, *10*(3), 241–255.

Rayner, K., Sereno, S. C., Morris, R. K., Schmauder, A. R., & Clifton Jr, C. (1989). Eye movements and on-line language comprehension processes. *Language and Cognitive Processes*, *4*(3-4), SI21–SI49.

Rayner, K., Well, A. D., & Pollatsek, A. (1980). Asymmetry of the effective visual field in reading. *Perception & Psychophysics*, *27*(6), 537–544.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological review*, *105*(1), 125.

Resnik, P. & Lin, J. (2010). Evaluation of NLP systems. *The handbook of computational linguistics and natural language processing*, *57*, 271.

Rohanian, O., Taslimipoor, S., Yaneva, V., & Ha, L. A. (2017). Using gaze data to predict multiword expressions. In *Proceedings of the international conference recent advances in natural language processing* (pp. 601–609).

San Agustin, J., Skovsgaard, H., Mollenbach, E., Barret, M., Tall, M., Hansen, D. W., & Hansen, J. P. (2010). Evaluation of a low-cost open-source gaze tracker. In *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 77–80). ACM.

Schmitt, N. & Underwood, G. (2004). Exploring the processing of formulaic sequences through a self-paced reading task. *Formulaic sequences: Acquisition, processing and use*, 173–189.

Seminck, O. & Amsili, P. (2018). A gold anaphora annotation layer on an eye movement corpus. In *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018)*.

Singh, A. D., Mehta, P., Husain, S., & Rajakrishnan, R. (2016). Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the workshop on computational linguistics for linguistic complexity* (pp. 202–212).

Siyanova-Chanturia, A. (2013). Eye-tracking and ERPs in multi-word expression research: a state-of-the-art review of the method and findings. *The Mental Lexicon*, *8*(2), 245–268.

Staub, A., Clifton Jr, C., & Frazier, L. (2006). Heavy np shift is the parser's last resort: evidence from eye movements. *Journal of memory and language*, *54*(3), 389–406.

Strzyz, M., Vilares, D., & Gómez-Rodríguez, C. (2019a). Towards making a dependency parser see. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 1500–1506).

Strzyz, M., Vilares, D., & Gómez-Rodríguez, C. (2019b). Viable dependency parsing as sequence labeling. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 717–723).

Stymne, S., Danielsson, H., Bremin, S., Hu, H., Karlsson, J., Lillkull, A. P., & Wester, M. (2012). Eye tracking as a tool for machine translation error analysis. In *Proceedings of the international conference on language resources and evaluation* (pp. 1121–1126).

Tokunaga, T., Nishikawa, H., & Iwakura, T. (2017). An eye-tracking study of named entity annotation. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 758–764.

Vainio, S., Hyönä, J., & Pajunen, A. (2003). Facilitatory and inhibitory effects of grammatical agreement: evidence from readers' eye fixation patterns. *Brain and Language*, *85*(2), 197–202.

Vasishth, S., von der Malsburg, T., & Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(2), 125–134.

Wallot, S., O'Brien, B., Coey, C. A., & Kelty-Stephen, D. (2015). Power-law fluctuations in eye movements predict text comprehension during connected text reading. In *Cognitive science* (pp. 2583–2588).

Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 233–243).

Williams, R. & Morris, R. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, *16*(1-2), 312–339.

Yaneva, V., Evans, R., Mitkov, R., et al. (2018). Classifying referential and non-referential it using gaze. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4896–4901).

Yaneva, V., Taslimipoor, S., Rohanian, O., et al. (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* (pp. 363–379). Springer.

Name	Reference	Text source	Annotations	Subjs.	Tokens	Coverage
GECO	Cop, Dirix, Drieghe, and Duyck (2017)	novel	named entities (Hollenstein and Zhang, 2019) multiword expressions (Rohanian, Taslimipoor, Yaneva, and Ha, 2017), referential uses of *it* (Yaneva, Evans, Mitkov, et al., 2018)	14	68606	95%
Dundee	Kennedy, Hill, and Pynte (2003)	newspaper articles	POS, Universal Dependencies (Barrett, Agić, and Søgaard, 2015), named entities (Hollenstein and Zhang, 2019), Anaphora (Seminck and Amsili, 2018)	10	58598	94%
CFILT-Sarcasm	Mishra, Kanojia, Nagar, Dey, and Bhattacharyya (2016a)	movie reviews	Sentiment	7	23466	85%
CFILT-Sentiment	Joshi, Mishra, Senthamilselvan, and Bhattacharyya (2014)	movie reviews	Sentiment	5	21076	82%
ZuCo	Hollenstein, Rotsztejn, et al. (2018)	Wiki/Sentiment	NER, Sentiment, Relations	12	13465	90%
CFILT-Scanpath	Mishra, Kanojia, Nagar, Dey, and Bhattacharyya (2017)	Wikipedia/simple Wikipedia	Reading difficulty scores	16	3677	89%
Provo	Luke and Christianson (2018)	miscellaneous domains	POS	470	2689	95%
CFILT-Coreference	Cheri, Mishra, and Bhattacharyya (2016)	MUC-6 dataset	Coreference	14	2210	95%

Table 1

Overview of eye tracking resources with normal, skilled reading for English by native speakers. Annotations column: annotations useful for NLP, if no reference is provided, the annotations are from the authors of original publication. Coverage = X% of vocabulary in corpus occurs in British National Corpus list[a] of most frequent English words (occurring more than 5 times). "CFILT" prefix denotes corpora from the Center for Indian Language Technology

[a]https://www.kilgarriff.co.uk/bnc-readme.html