Sequence labelling and classification with gaze: Novel uses of eye-tracking data for natural language processing

Abstract

Eye-tracking data 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. It has been studied extensively in order to understand human cognition but has only recently been considered for natural language processing (NLP). This article provides a comprehensive overview of how gaze data is being used in data-driven NLP, particularly for sequence labelling and sequence classification tasks, and based on this argues that it may effectively counter one of the core challenges of machine-learning-based NLP: the scarcity of annotated data. The recent advances in NLP using gaze are used to discuss how the gaze signal from human readers can be leveraged and the potentials and limitations of the data source are also considered. The article also provides an overview of the largest, English eye tracking corpora of naturalistic reading, which are usable for NLP.

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

Sequence labelling and 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 psycho-linguistic studies, word-based eye movements metrics have proven to be strongly correlated with high-level text processing, such as syntactic and semantic structures (Rayner et al., 1989).

Natural language processing (NLP) tries to solve tasks within sequence labelling and sequence classification, which are implicitly performed by skilled human readers, e.g. sentiment classification, syntactic analysis, and the eye movements reflect these cognitive processes. 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). The 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 article first introduces 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 with respect to sequence labelling and sequence classification using gaze data. Finally, we will summarise observations across the studies in 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 article is data-driven NLP: namely sequence labelling and sequence classification tasks on English text using eye movements from adults performing *naturalistic reading*. Naturalistic reading denotes self-paced reading of naturally-occurring text without any task solving[1] or reading constraints, such as

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

limiting the preview of the following word. We, therefore, do not cover the fairly large body of work on annotators' eye movements but do include Joshi et al. (2014) and Mishra et al. (2016a) where the annotation process is separated from the reading and gaze recording. Mishra and Bhattacharyya (2018) presented a review on the cognitive effort of annotation.

In the main parts of the survey, we focus on studies where an actual *evaluation* of an NLP sequence labelling or sequence classification task has taken place but refer to experimental or large-scale psycho-linguistic studies that *describe* correlations between eye movements and linguistic phenomena to introduce each section. We do not include dialogue, studies modelling reader attributes from eye movements, nor information retrieval studies.

Very Brief Introduction to Eye Movements

Contrary to the perceived experience of reading, the eyes do not glide smoothly across the lines of text. Instead, the eyes interchangeably fixate and perform rapid, ballistic movements; saccades. Fixations last averagely around 200-250ms, with large variations. Saccades last around 20-40ms and move the eyes 7-9 letter spaces forward. The eye does not take in information during saccades. In normal reading, around 10% of the saccades – often unconscious to the reader – move back to an earlier read part of the text for further processing, maybe even re-processing. This is called *regressions*. Eye movements thereby allow us to study early and late cognitive processing separately via a range of established, word-based metrics making this one of the richest known data sources on human text processing during reading.

It is well-established that fixation durations on a word are sensitive to a wide range of lexical and linguistic properties of a word. Rayner (1998) provides a review. There is a tight relationship between the cognitive processing of a word and the fixation duration on that word during reading. *Spill-over* and *preview* effects are phenomena that demonstrate that this is not an air-tight relationship. Spill-over effects occur when a property of a word elicits longer fixation durations on the subsequently fixated word.

Word frequency is known to cause spill-over effects (Rayner & Duffy, 1986). Preview effects and its implications are discussed below.

It may 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, e.g., syntactic and semantic processing, which are of interest to NLP. NLP Models should, however, include low-level processing features because they interact with high-level processing.

Low-level Processing

Due to the anatomy of the eye, readers can only see a small part of the text during each fixation. It is an asymmetric area that, for a skilled reader of an alphabetic language arranged left-to-right, extends 3-4 characters to the left of the fixation and 14-15 characters to the right (Rayner et al., 1980). In practice, words can only be identified in an even smaller area only extending 7-8 characters to the right of the fixation, called *the perceptual span* (McConkie & Rayner, 1976). The saccadic length and the size of the perceptual span vary as a function of text difficulty and reading skill level.

It is a robust finding that fixations are shorter if the reader can get a preview of the word. This is because some processing starts before fixating word. Some words are even processed sufficiently during preview to skip them due to a complex interaction of e.g., orthography, frequency, and predictability (Drieghe et al., 2005; Starr & Inhoff, 2004; White, 2008).

High-level processing

Fixation durations are shorter if a word is easy to identify and understand (Clifton et al., 2007). "To be easy to identify and understand" covers a range of high-level effects. One of the most studied effects on fixation duration is word frequency. It is consistent that readers look longer at infrequent words and/or long words (Just & Carpenter, 1980; Rayner, 1977). Just and Carpenter (1980) found that almost 70% of the variance in mean gaze duration was explained by word length and word frequency. But gaze

duration is also sensitive to a range of other factors, e.g., predictability from context Inhoff (1984)), age of acquisition (Kemper & Liu, 2007), familiarity (Juhasz & Rayner, 2003; Williams & Morris, 2004), and morphology (Hyönä et al., 2004). In each section of the survey, we will cover psycho-linguistic evidence of relevant high-level effects.

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

When not in a controlled setup, a challenge about eye movements is that they reflect *a* cognitive process, but not *which*, meaning that 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

Text stimulus from psycho-linguistic studies often consists of constructed sentences that have very low frequency in naturally-occurring text. The sentences are often read out of context or reading is constrained. This makes this data difficult to use for NLP purposes.

Table 1 provides an overview over available, larger (>2000 words), English corpora of naturalistic reading of naturally-occurring text by native readers. Some of the corpora such as the Dundee Corpus (Kennedy et al., 2003) and the GECO corpus (Cop et al., 2017) also include parts in another language/read by non-native subjects. These parts are *not* included in the table.

A Survey on using gaze for Sequence Labelling and Sequence Classification

This comprehensive survey is thematically ordered according to NLP topic where relevant work on sequence labelling and sequence classification tasks using gaze data has been carried out. Each section is introduced with a summary of psycho-linguistic findings – both experimental and on larger quantities of naturalistic reading data –

concerning each topic. Due to space limitations, we can not provide background and context for each NLP task, but rely on the reader to know or seek this information herself.

Text Complexity

Text comprehension has been thoroughly studied through eye movements, e.g., (Rayner et al., 2006; Vasishth et al., 2013). Processing difficulties show effects on regressions, saccade length, and fixation durations. Štajner et al. (2017) showed that gaze metrics correlate with text complexity metrics. It is the psycho-linguistic studies on text comprehension that motivate using gaze for text complexity evaluation in NLP. In NLP, gaze has been used for scoring the text complexity when evaluating the readability, grammaticality or acceptability on its own or in the context of evaluating the output from automatic systems.

González-Garduño and Søgaard (2017) showed that text readability classification worked better when using gaze as an auxiliary task in a multi-task learning setup and Singh et al. (2016) used predicted gaze metrics learned from human reading to predict the readability. Mathias et al. (2018) improved text quality evaluation using gaze.

Evaluation of Machine Output. It is common to use human evaluation as a reference when exploring how to best evaluate the output from automatic systems (Chaganty et al., 2018; Resnik & Lin, 2010). But eye movements during reading has also been used as a signal to score automatically generated text. Gaze can, in this context, be considered noisy annotations. On smaller data sets, there are results suggesting that eye movements reveal the location of machine translation errors (Bremin et al., 2010; Doherty & O'Brien, 2009) and gaze metrics may even help to distinguish different types of errors (Stymne et al., 2012). Klerke et al. (2015b) found that reading metrics were better proxies for the usability of translated text than the standard, automatic metric, bilingual evaluation understudy score (Papineni et al., 2002).

Klerke et al. (2015a) used gaze to evaluate the output of automatic sentence compression. Due to this correlation, gaze features have also been used to improve

sentence compression in a multi-task setup where gaze features were predicted as an auxiliary task to improve sentence compression (Klerke et al., 2016).

Scan Path Metrics. The scan path is the trajectory of the eyes through the text. Apart from the word-based metrics, which are mainly used, the scan path over a text sequence has also been summarised in quantitative metrics, mainly for text comprehension. von der Malsburg and Vasishth (2011) presented a sentence-level score, *Scasim*, that was used to detect irregularities in the scan path during reading. Mishra et al. (2017b) presented another scan path metric, *scan path complexity*, and showed that it 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 et al. (2015) also showed that the degree of power-law scaling in raw eye movements was predictive of text comprehension.

Evaluation of other Resources

Apart from evaluating the output from automatic systems, such as translation and compression, eye movements have also been used to evaluate word embeddings and sentence representations. Though this is not in itself a sequence labelling or classification task, this is related because such representations of text are common building blocks of systems performing these tasks.

Frank (2017) and Søgaard (2016) evaluated word embeddings with gaze data from reading with contradictory results. On a similar note, Abdou et al. (in review) showed that there is a strong, positive correlation between gaze metrics and disagreements between ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) sentence representations. Here, gaze serves as a triangulation metric validating the disagreement of language encoders.

Part-of-speech

Carpenter and Just (1983) noted that 38% of function words are fixated and 83% of content words are fixated. Bauman (2013), Demberg and Keller (2008) and Pynte and Kennedy (2007) found a negative correlation between the word class probability

and fixation probability/duration. 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 (Pynte et al., 2009). None of 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 POS processing is dependent on the context.

There are, however, results from the NLP, that suggest that eye movements can disambiguate some word classes (Barrett & Søgaard, 2015a). They also show that word-based gaze features can be used to improve supervised, type-constrained POS tagging. Word-based eye movement features have also been used to significantly improve type-constrained, weakly supervised POS induction (Barrett et al., 2016a; Barrett et al., 2018b). Barrett et al. (2016b) even suggest that correlations may transfer across the related languages English and French. Here, English gaze data can be used to improve POS induction for French.

Syntax

In their exhaustive survey over higher-level effects in psycho-linguistic studies of human reading, Clifton et al. (2007) found that all four studies on syntactic complexity in sentences without syntactic ambiguity showed an effect of syntactic complexity on early gaze measures. (Hyönä & Vainio, 2001; Rayner et al., 1989; Staub et al., 2006; Vainio et al., 2003). This was also backed up 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 durations for nouns. It is however more controversial how exactly the human parser works. Out of the 100 studies on syntax, semantics and pragmatics in Clifton et al. (2007), 70 explored temporarily ambiguous syntax/garden path sentences, which is an interesting phenomenon when the objective is understanding the human syntax parser. There is evidence that humans build the syntactic structure of a sentence incrementally during reading in a (partially) top-down manner. This is opposed to a bottom-up parser, which

adds nodes to the phrase marker on the basis of encountered input and enters a higher level node only after some or all of the node's daughters have already been encountered.

From the field of NLP, Barrett and Søgaard (2015b) show that eye movements can, to some extent, disambiguate four syntactic roles for nouns. They also demonstrate that eye movements help supervised dependency parsing better than pre-trained word embeddings when using little training data. Combining eye movement features with other measures reflecting cognitive text processing helps weakly supervised, type-constrained, syntactic chunking better than pre-trained word embeddings (Barrett et al., 2018b). Klerke and Plank (in review) also find that predicting a gaze feature as an auxiliary task may help chunk boundary and POS tagging a multitask learning setup.

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 is to learn aspects of the text. 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 showing that only in some cases, is irony processed slower (Filik et al., 2014; Gibbs, 1986; Gibbs Jr, 1994; Ivanko & Pexman, 2003).

Mishra et al. (2017a), Mishra et al. (2016a), Mishra et al. (2016b) showed that eye movements helped to predict whether a reader caught the sarcastic meaning of a sentence or not. Joshi et al. (2015) also used implicit and explicit context features along with the gaze features.

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): As already mentioned, word familiarity and predictability has a negative effect on fixation duration. Additionally, reading patterns are a reliable indicator 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) should have a distinct reading pattern. Tokunaga et al. (2017) analysed eye tracking signals during the annotation of named entities to find useful features for NER. Their work shows that humans take into account a broad context to identify named entities, including predicate-argument structure, which hints to the usefulness of eye tracking recordings of full sentences.

Hollenstein and Zhang (2019) found improvements when augmenting NER models with eye tracking data. It also seems like gaze may be used to express relations in text: results from the statistical analysis of Jaffe et al. (2018) suggested that gaze might help co-reference resolution and Yaneva et al. (2018) used eye movements to classify referential and non-referential uses of *it*.

Multiword Expressions

Multiword expressions (MWEs) vary in their linguistic properties but they are perceived as highly conventional by L1 speakers (Siyanova-Chanturia, 2013). MWEs are an example of eye movement processing on the super-word level. In behavioural experiments using including eye tracking, the entire MWE is found to have a processing advantage over novel strings of language (Schmitt & Underwood, 2004; Yaneva et al., 2017). Rohanian et al. (2017) found that MWEs could be predicted from gaze features.

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 et al., 2016c), but eye tracking features could be learned automatically from scan paths (Mishra et al., 2017a). In a different NLP study, Hollenstein et al. (2019) showed improvements not only on binary (*positive/negative*) but also on ternary (*positive/negative/neutral*) sentiment analysis when using gaze. Mishra and Bhattacharyya (2018) presented a more detailed overview of extracting gaze features for sentiment analysis. Barrett et al. (2018a) used gaze to regularise the

attention function for sentiment classification as well as for detection of grammatical error and hate speech and Long et al. (2019) used predicted total gaze time to build an attention layer to improve sentiment classification.

How to use gaze for NLP

This section contains a discussion based on observations across NLP studies from the survey above.

Which Gaze Features?

Some NLP studies systematically tried different groups of gaze features in order to identify the best gaze feature group for the task and many found that the best feature combination is using all the gaze features. This is the case for supervised MWE (Rohanian et al., 2017), NER (Hollenstein & Zhang, 2019) as well as POS (Barrett et al., 2016a). When studying the contribution of individual gaze features, Barrett and Søgaard (2015a, 2015b) and Mishra et al. (2016c) found that the signal is distributed over many eye movement features for classification of POS, grammatical function, sentiment, and sarcasm, respectively. The right eye tracking features depend on the task, but for these complex phenomena, the cognitive processing seems to be distributed over a wide range of word-level, eye-tracking features, that spans both early and late processing measures.

Many studies also combined eye movement features with other features that either supplement the features or are considered confounding factors and find that this works better for machine learning algorithms that eye movement alone: Mishra et al. (2016c) systematically combined eye movement features with sentiment features, sarcasm, irony and thwarting related features, and textual features related to reading difficulty and find that combining all feature groups overall is better for predicting sarcasm and sentiment. Yaneva et al. (2018) also found than eye movements combined with linguistic features work better for classifying referential uses of *it* than models based only on linguistic features and models based on only on gaze. The best model in Barrett et al. (2018b) for POS induction used pre-trained word embeddings combined

with eye movement features. Barrett and Søgaard (2015a, 2015b) showed that including word length and word frequencies were better than using gaze alone for POS tagging and dependency parsing. The best model in Rohanian et al. (2017) combined both gaze feature groups with POS and frequency features.

Mishra et al. (2017a) presented an interesting approach where the gaze representation is learned in a convolutional neural network from the raw scan path instead of, as all the remaining studies in the survey, relying on manually selected features. This approach yielded better performance for sentiment and sarcasm detection than using hand-crafted features.

Gaze Data at Test Time and How to Include Features

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

The simplest approach is to include gaze features as multi-dimensional vectors to representing each word, possibly along with other word-based features, as done by e.g. Barrett and Søgaard (2015a), Rohanian et al. (2017) and Yaneva et al. (2018). However, this requires gaze data at test time. But Barrett et al. (2016a), Barrett et al. (2016b) showed that word-type averages of gaze features helped POS induction better than token-level features, thereby using gaze representation similarly to word embeddings with which they have also been combined (Barrett et al., 2018b). Klerke and Plank (in review) showed that word type variance was better than individual gaze representations and less aggregated gaze features. Type-level gaze features do not require gaze-annotated test data.

Singh et al. (2016) introduced a method where eye movements are learned in order to alleviate the need to get the task data annotated with eye movements. A similar approach is also used by Long et al. (2019).

González-Garduño and Søgaard (2017), Klerke et al. (2016) and Klerke and Plank

(in review) employ a multi-task learning setup for text compression, readability prediction, and syntactic tagging, respectively, while also learning to predict a gaze feature. This approach also did not need gaze-annotating the main task test set. Another option is to regularise the attention of a recurrent neural network with gaze for sequence classification. Attention weights the influence that each word has on the model, but requires lots of data to learn during normal training. Barrett et al. (2018a) used sentences from the main dataset to update the model parameters, while sentences from a smaller, non-overlapping eye tracking corpus were used to only train the attention function.

The Case for and Against Aggregation

Controlled psycho-linguistic studies include enough readers to obtain significant differences considering the effect sizes of interest (Vasishth et al., 2018). Many NLP studies that use eye movements as word representations, averaged the eye movement metrics over several users arguing for more stability and less noise, but most studies are limited by the corpora (Hollenstein et al., 2019; Mishra et al., 2016c; Rohanian et al., 2017; Yaneva et al., 2018). But how many readers should one average over to obtain a robust signal for NLP? Gaze annotation can never be a gold annotation, irrespective of the number of readers. It is intrinsically noisy and there is no correct reading pattern. But skilled readers will exhibit a more idiosyncratic reading behaviour under similar condition. Language learners or readers with reading impairments will exhibit a noisier signal, that is difficult to use in NLP (Bingel et al., 2018).

Hollenstein and Zhang (2019) used eye movement features to improve named entity recognition, relation classification and sentiment classification and showed that averaging over ten skilled readers is able to swallow the noise to the extent where the average worked almost as good as the best individual reader.

Potentials

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 few large resources of gaze-annotated data 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[2], and only a few of them have annotated resources to train supervised models (Plank, 2016b). It is faster and cheaper to have skilled, native readers read a text than professional annotators to annotate it. For some low-resource languages, trained annotators may be hard to find. Moreover, eye trackers are expected to be standard equipment in most near-future consumer hardware, which entails the availability of larger quantities of eye tracking data. This survey contains evidence that eye movements from skilled readers contain traces of human cognitive processing of linguistic phenomena that machines struggle to learn. The signal can be leveraged by unsupervised algorithms; alone (Barrett et al., 2016a; Barrett et al., 2016b) or combined with word embeddings or other accessible human text processing features (Barrett et al., 2018b).

Other Human Data Sources

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

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 affordable and available data source. Advancing the text one word at the time by keypress does not reflect the intricate processes of natural reading. It prevents preview effects and does not measure regressions, but predicting crowdsourced reading times as an auxiliary task in a multi-task learning set up drastically improved syntactic chunking, CCG supertagging,

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

and detection of MWE and outperformed gaze. Christensen et al. (in review).

Other data sources also provide indirect information about human cognitive text processing on word-level. Reading times, decision times and keystroke metrics have been used to improve and evaluate NLP tasks. This data can be collected with current consumer technology. For instance, Auguste et al., 2017 evaluated word embeddings against reading times in lexical decision tasks. (Plank, 2016a) used keystroke logs to aid shallow parsing. Furthermore, acoustic cues have been also used for parsing (Pate & Goldwater, 2013) and chunking (Pate & Goldwater, 2011). When combining word-type averages of multiple data sources reflecting human cognitive processing, eye movements and prosodic features were significantly better than word embeddings alone for unsupervised chunking (Barrett et al., 2018b).

Also, 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 et al., 2018) provides simultaneous eye tracking and EEG recordings, which have been leveraged to improve various information extraction tasks (Hollenstein et al., 2019). However, EEG is a very noisy data a source. Unlike eye tracking, it captures much more of the cognitive processing load than merely the reading process. While EEG did yield improvement in those tasks, eye tracking still shows better results.

fMRI data has been used more widely than EEG: Bingel et al. (2016) extracted token-level signals of syntactic processing from fMRI for POS induction; Wehbe et al. (2014) recorded data from subjects reading stories and aligned statistical language models with brain activity; and fMRI data has also been used to evaluate word embeddings (Abnar et al., 2018; Søgaard, 2016).

Neuroimaging data sources are noisier, more expensive and more cumbersome to acquire. Moreover, while eye tracking and EEG 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, are also in their relative effortless method of recording and their higher signal-to-noise ratio.

Data Acquisition

In the past decade, eye trackers have evolved from expensive laboratory equipment to being built into consumer hardware. Eye trackers come as do-it-yourself kits (Krafka et al., 2016; San Agustin et al., 2010) using inexpensive, consumer-grade hardware or webcams from mobile devices (Papoutsaki et al., 2016). Samsung released its first cell phone with an eye tracker in 2013 and shortly after, the first $100 eye tracker was released. Since then, more low-cost hardware options have emerged, also from established eye tracker companies. iPhone X comes with API access to all necessary components for external developers to implement eye tracking.

Conclusion

Humans implicitly perform many of the tasks that NLP systems try to learn. The cognitive processing of text has been studied carefully via eye movements. We presented an overview of current NLP applications for sequence labelling and sequence classification tasks harnessing human cognition through the use of eye movement features. The usages span over a range of NLP tasks: semantic, syntactic, POS, and relational.

The increasing availability of eye tracking data should further encourage the usage of gaze data in more NLP and will allow for researching additional methods 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

Abdou, M., Kulmizev, A. & Søgaard, A. (in review). Analysing the representational geometry of neural language encoders.

Abnar, S., Ahmed, R., Mijnheer, M. & Zuidema, W. (2018). Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, 57–66.

Auguste, J., Rey, A. & Favre, B. (2017). Evaluation of word embeddings against cognitive processes: Primed reaction times in lexical decision and naming tasks. *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, 21–26

Evaluate word embeddings through lexical decision times.

Barrett, M., Agić, Ž. & Søgaard, A. (2015). The Dundee Treebank. *The 14th International Workshop on Treebanks and Linguistic Theories*.

Barrett, M., Bingel, J., Hollenstein, N., Rei, M. & Søgaard, A. (2018a). Sequence classification with human attention. *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 302–312.

Barrett, M., Bingel, J., Keller, F. & Søgaard, A. (2016a). Weakly supervised part-of-speech tagging using eye-tracking data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, *2*, 579–584.

Barrett, M., González-Garduño, A. V., Frermann, L. & Søgaard, A. (2018b). Unsupervised induction of linguistic categories with records of reading, speaking, and writing. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 2028–2038.

Barrett, M., Keller, F. & Søgaard, A. (2016b). Cross-lingual transfer of correlations between parts of speech and gaze features. *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, 1330–1339.

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

Barrett, M. & Søgaard, A. (2015b). Using reading behavior to predict grammatical functions. *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, 1–5.

Bauman, P. (2013). Syntactic category disambiguation within an architecture of human language processing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*(35), 1833–1838.

Bingel, J., Barrett, M. & Klerke, S. (2018). Predicting misreadings from gaze in children with reading difficulties. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 24–34.

Bingel, J., Barrett, M. & Søgaard, A. (2016). Extracting token-level signals of syntactic processing from fMRI - with an application to PoS induction. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, *1*, 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. (1983). What your eyes do while your mind is reading. *Eye movements in reading: Perceptual and language processes*, 275–307.

Chaganty, A., Mussmann, S. & Liang, P. (2018). The price of debiasing automatic metrics in natural language evalaution. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, *1*, 643–653.

Cheri, J., Mishra, A. & Bhattacharyya, P. (2016). Leveraging annotators' gaze behaviour for coreference resolution. *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, 22–26.

Christensen, N. L., Netterstrøm, R., Barrett, M. & Søgaard, A. (in review). Leveraging syntactic biases in self-paced reading.

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

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

Drieghe, D., Rayner, K. & Pollatsek, A. (2005). Eye movements and word skipping during reading revisited. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(5), 954.

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.

Frank, S. L. (2017). Word embedding distance does not predict word reading time.

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

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

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. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 438–443.

Grice. (1975). Logic and conversation. *Speech acts: Syntax and semantics* (pp. 41–58). Academic Press.

Hollenstein, N., Barrett, M., Troendle, M., Bigiolli, F., Langer, N. & Zhang, C. (2019). Advancing NLP with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*.

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. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, 1–9.

Joshi, A., Mishra, A., Senthamilselvan, N. & Bhattacharyya, P. (2014). Measuring sentiment annotation complexity of text. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, *2*, 36–41.

Joshi, A., Sharma, V. & Bhattacharyya, P. (2015). Harnessing context incongruity for sarcasm detection. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, *2*, 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. & 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. *Proceedings of the 12th European conference on eye movement*.

Klerke, S., Alonso, H. M. & Søgaard, A. (2015a). Looking hard: Eye tracking for detecting grammaticality of automatically compressed sentences. *Proceedings of the 20th Nordic Conference of Computational Linguistics*, 97–105.

Klerke, S., Castilho, S., Barrett, M. & Søgaard, A. (2015b). Reading metrics for estimating task efficiency with MT output. *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, 6–13.

Klerke, S., Goldberg, Y. & Søgaard, A. (2016). Improving sentence compression by learning to predict gaze. *Proceedings of 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 1528–1533.

Klerke, S. & Plank, B. (in review). At a glance: Gaze aggregation impact on syntactic tagging.

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

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. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, *1*, 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. (2017a). Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, *1*, 377–387.

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

Mishra, A., Kanojia, D., Nagar, S., Dey, K. & Bhattacharyya, P. (2016b). Harnessing cognitive features for sarcasm detection. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, *1*, 1095–1104.

Mishra, A., Kanojia, D., Nagar, S., Dey, K. & Bhattacharyya, P. (2016c). Leveraging cognitive features for sentiment analysis. *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 156–166.

Mishra, A., Kanojia, D., Nagar, S., Dey, K. & Bhattacharyya, P. (2017b). Scanpath complexity: Modeling reading effort using gaze information. *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. *Proceedings of the 40th annual meeting on Association for Computational Linguistics*, 311–318.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J. & Hays, J. (2016). WebGazer: Scalable webcam eye tracking using user interactions. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*.

Pate, J. K. & Goldwater, S. (2011). Unsupervised syntactic chunking with acoustic cues: Computational models for prosodic bootstrapping. *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 20–29.

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

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2227–2237.

Plank, B. (2016a). Keystroke dynamics as signal for shallow syntactic parsing. *Proceedings of the 25th International Conference on Computational Linguistics*, 609–618.

Plank, B. (2016b). What to do about non-standard (or non-canonical) language in NLP. *Proceedings of the Conference on Natural Language Processing (KONVENS)*, 13–20.

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.

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

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

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*(3), 372–422.

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

Rayner, K., 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.

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. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 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. *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, 77–80.

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.

Singh, A. D., Mehta, P., Husain, S. & Rajakrishnan, R. (2016). Quantifying sentence complexity based on eye-tracking measures. *Proceedings of the workshop on Computational Linguistics for Linguistic Complexity*, 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.

Søgaard, A. (2016). Evaluating word embeddings with fMRI and eye-tracking. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 116–121.

Štajner, S., Yaneva, V., Mitkov, R. & Ponzetto, S. P. (2017). Effects of lexical properties on viewing time per word in autistic and neurotypical readers. *Proceedings of the 12th Workshop on Innovative Use ofNLP for Building Educational Applications*, 271–281.

Starr, M. & Inhoff, A. (2004). Attention allocation to the right and left of a fixated word: Use of orthographic information from multiple words during reading. *European Journal of Cognitive Psychology*, *16*(1-2), 203–225.

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.

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. *Proceedings of the International Conference on Language Resources and Evaluation*, 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., Mertzen, D., Jäger, L. A. & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175.

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. *Cognitive Science*, 2583–2588.

Wehbe, L., Vaswani, A., Knight, K. & Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 233–243.

White, S. J. (2008). Eye movement control during reading: Effects of word frequency and orthographic familiarity. *Journal of experimental psychology: Human perception and performance*, *34*(1), 205.

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

Yaneva, V., Evans, R., Mitkov, R. et al. (2018). Classifying referential and non-referential it using gaze. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 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. *International conference on computational and corpus-based phraseology*, 363–379.

Name	Reference	Text source	Annotations	Subjs.	Tokens	Coverage
GECO	Cop et al. (2017)	novel	NER (Hollenstein and Zhang, 2019) MWE (Rohanian et al., 2017), referential uses of *it* (Yaneva et al., 2018)	14	68606	95%
Dundee	Kennedy et al. (2003)	newspaper articles	POS, Universal Dependencies (Barrett et al., 2015), NER (Hollenstein and Zhang, 2019)	10	58598	94%
CFILT-Sarcasm	Mishra et al. (2016b)	movie reviews	Sentiment	7	23466	85%
CFILT-Sentiment	Joshi et al. (2014)	movie reviews	Sentiment	5	21076	82%
ZuCo	Hollenstein et al. (2018)	Wiki/Sentiment	NER, Sentiment, Relations	12	13465	90%
CFILT-Scanpath	Mishra et al. (2017b)	Wikipedia/simple Wikipedia	Reading difficulty scores	16	3677	89%
Provo	Luke and Christianson (2018)	miscellaneous domains	POS	470	2689	95%
CFILT-Coreference	Cheri et al. (2016)	MUC-6 dataset	Coreference	14	2210	95%

Table 1

Overview over 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 (occuring more than 5 times). "CFILT" prefix denotes corpora from the Center for Indian Language Technology

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