Using reading behavior to predict grammatical functions

Maria Barrett and Anders Søgaard
University of Copenhagen
Njalsgade 140
DK-2300 Copenhagen S
{barrett,soegaard}@hum.ku.dk

Abstract

This paper investigates to what extent grammatical functions of a word can be predicted from gaze features obtained using eye-tracking. A recent study showed that reading behavior can be used to predict coarse-grained part of speech, but we go beyond this, and show that gaze features can also be used to make more fine-grained distinctions between grammatical functions, e.g., subjects and objects. In addition, we show that gaze features can be used to improve a discriminative transition-based dependency parser.

1 Introduction

Readers fixate more and longer on open syntactic categories (verbs, nouns, adjectives) than on closed class items like prepositions and conjunctions (Rayner and Duffy, 1988; Nilsson and Nivre, 2009). Recently, Barrett and Søgaard (2015) presented evidence that gaze features can be used to discriminate between most pairs of parts of speech (POS). Their study uses all the coarse-grained POS labels proposed by Petrov et al. (2011). This paper investigates to what extent gaze data can also be used to predict grammatical functions such as subjects and objects. We first show that a simple logistic regression classifier trained on a very small seed of data using gaze features discriminates between some pairs of grammatical functions. We show that the same kind of classifier distinguishes well between the four main grammatical functions of nouns, POBJ, DOBJ, NN and NSUBJ. In §3, we also show how gaze features can be used to improve dependency parsing. Many gaze features correlate with word length and word frequency (Rayner, 1998) and these could be as good as gaze features, while being easier to obtain. We use frequencies from the unlabelled portions of the English Web Treebank and word length as baseline in all types of experiments and find that gaze features to be better predictors for the noun experiment as well as for improving parsers.

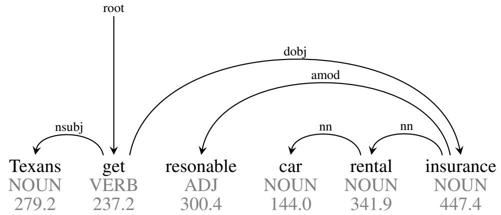


Figure 1: A dependency structure with average fixation duration per word

This work is of psycholinguistic interest, but we show that gaze features may have practical relevance, by demonstrating that they can be used to improve a dependency parser. Eye-tracking data becomes more readily available with the emergence of eye trackers in mainstream consumer products (San Agustin et al., 2010). With the development of robust eye-tracking in laptops, it is easy to imagine digital text providers storing gaze data, which could then be used as partial annotation of their publications.

Contributions We demonstrate that we can discriminate between some grammatical functions using gaze features and which features are fit for the task. We show a practical use for data reflecting human cognitive processing. Finally, we use gaze features to improve a transition-based dependency parser, comparing also to dependency parsers augmented with word embeddings.

Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning, pages 1–5,
Lisbon, Portugal, 18 September 2015. ©2015 Association for Computational Linguistics.

2 Eye tracking data

The data comes from (Barrett and Søgaard, 2015) and is publicly available[1]. In this experiment 10 native English speakers read 250 syntactically annotated sentences in English (min. 3 tokens, max. 120 characters). The sentences were randomly sampled from one of five different, manually annotated corpora from different domains: Wall Street Journal articles (WSJ), Wall Street Journal headlines (HDL), emails (MAI), weblogs (WBL), and Twitter (TWI)[2]. See Figure 1 for an example.

Features It is not yet established which eye movement reading features are fit for the task of distinguishing grammatical functions of the words. To explore this, we extracted a broad selection of word- and sentence-based features. The features are inspired by Salojärvi et al. (2003) who used a similar exploratory approach. For a full list of features, see Appendix.

2.1 Learning experiments

In our binary experiments, we use L2-regularized logistic regression classifiers with the default parameter setting in SciKit Learn[3] and a publicly available transition-based dependency parser[4] trained using structured perceptron (Collins, 2002; Zhang and Nivre, 2011).

Binary classification We trained logistic regression models to discriminate between pairs of the 11 most frequent dependency relations where the sample size is above 100: (AMOD, NN, AUX, PREP, NSUBJ, ADVMOD, DEP, DET, DOBJ, POBJ, ROOT) only using gaze features. E.g., we selected all words annotated as PREP or NSUBJ and trained a logistic regression model to discriminate between the two in a five-fold cross validation setup. Our baseline uses the following features: word length, position in sentence and word frequency.

Some dependency relations are almost uniquely associated with one POS, e.g. determiners where 84.8% of words with the dependency relation DET are labeled determiners. This means that in some cases, the grammatical function of a word follows from its part of speech. In another binary experiment, we therefore focus on nouns to show that eye movements *do* make more fine-grained distinctions between different grammatical functions. Nouns are mostly four-way ambiguous: 74.6% of the 946 nouns in the dataset have one of four dependency relations to its head. Nouns with POBJ relations is 18.9% of all nouns, NSUBJ is 17.0%, NN is 27.0% and DOBJ is 14.9%. The remaining 25.4% of the nouns are discarded from the noun experiment since they have 28 different relations to their head.

RANK	FEATURE NAME	% OF VOTES
0	Next word fixation probability	13.46
1	Fixation probability	11.14
2	n Fixations	9.66
3	Probability to get 2nd fixation	8.90
4	Previous word fixation probability	7.17
5	n Regressions from	5.65
6	First fixation duration on every word	5.45
7	Mean fixation duration per word	5.17
8	Previous fixation duration	4.93
9	Re-read probability	4.65
10	Probability to get 1st fixation	4.53
11	n Long regressions from word	3.77
12	Share of fixated words per sent	3.04
13	n Re-fixations	1.88
14	n Regressions to word	1.76

Table 1: Most predictive features for binary classification of 11 most frequent dependency relations using five-fold cross validation.

Parsing In all experiments we trained our parsing models on four domains and evaluated on the fifth to avoid over-fitting to the characteristics of a specific domain. All parameters were tuned on the WSJ dataset. We did 30 passes over the data and used the feature model in Zhang and Nivre (2011) – concatenated with gaze vectors for the first token on the buffer, the first token in the stack, and the left sibling of the first token in the stack. We extend the feature representation of each parser configuration by 3×26 features. Our gaze vectors were normalized using the technique in Turian et al. (2010) ($\sigma \cdot E/SD(E)$) using a scaling factor of $\sigma = 0.001$. Gaze features such as fixation duration are known to correlate with word frequency and word length. To investigate whether word length and frequency are stronger features than gaze, we perform an experiment, +FREQ+LEN, where our baseline and system also use frequencies and word length as features.

[1] https://bitbucket.org/lowlands/release/src

[2] Wall Street Journal sentences are from OntoNotes 4.0 release of the English Penn Treebank. catalog.ldc.upenn.edu/LDC2011T03. Mail and weblog sentences come from the English Web Treebank. catalog.ldc.upenn.edu/LDC2012T13. Twitter sentences are from the work of (Foster et al., 2011)

[3] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[4] https://github.com/andersjo/hanstholm

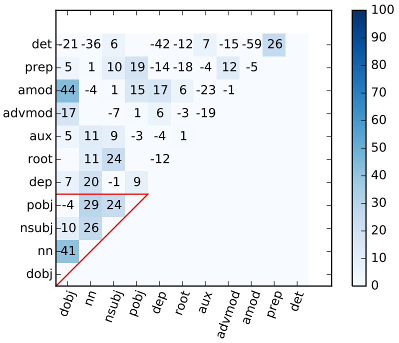


Figure 2: Error reduction over the baseline for binary classifications of 11 most frequent dependency relations. 5-fold cross validation. Dependency relations associated with nouns in triangle.

3 Results

Predictive features To investigate which gaze features were more predictive of grammatical function, we used stability selection (Meinshausen and Bühlmann, 2010) with logistic regression classification on binary dependency relation classifications on the most frequent dependency relations.

For each pair of dependencies, we perform a five-fold cross validation and record the informative features from each run. Table 1 shows the 15 most used features in ranked order with their proportion of all votes. The features predictive of grammatical functions are similar to the features that were found to be predictive of POS (Barrett and Søgaard, 2015), however, the probability that a word gets first and second fixation were not important features for POS classification, whereas they are contributing to dependency classification. This could suggest that words with certain grammatical functions are consistently more likely or less likely to get first and second fixation, but could also be due to a frequent syntactic order in the sample.

Binary discrimination Error reduction over the baseline can be seen in Figure 2. The mean accuracy using logistic regression on all binary classification problems between grammatical functions is 0.722. The frequency-position-word length baseline is 0.706. In other words, using gaze features leads to a 5.6% error reduction over the baseline. The worst performance (where our baseline outperforms using gaze features) is seen where one relation is associated with closed class words (DET, PREP, AUX), and where discrimination is easier.

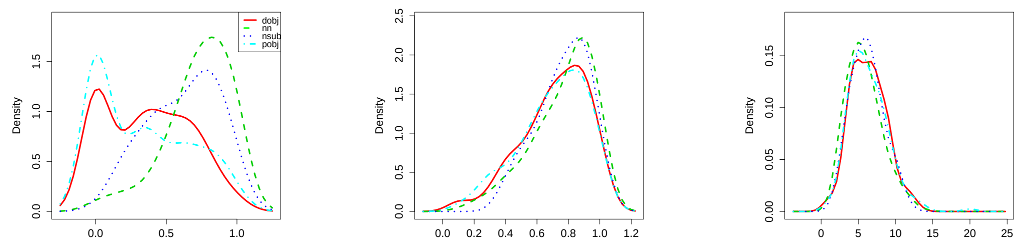
RANK	FEATURE NAME	% OF VOTES
0	Next word fixation probability	20.66
1	Probability to get 2nd fixation	19.83
2	nRegressions from word	14.05
3	Previous word fixation probability	8.68
4	Probability to get 1st fixation	7.44

Table 2: Most predictive features for the binary classification of four most frequent dependency relations for nouns using five-fold cross validation.

Noun experiment Error reductions for pairwise classification of nouns are between -4% and 41%. See Figure 2. The average accuracy for binary noun experiments is 0.721. Baseline accuracy is 0.647. For POBJ and DOBJ the baseline was better than using gaze, but for the other pairs, gaze was better. When doing stability selection for nouns with only the four most frequent grammatical functions, the most important features can be seen from Figure 2. The most informative feature is the fixation probability of the next word. Kernel density of this feature can be seen in Figure 3a, and it shows two types of behavior: POBJ and DOBJ, where the next word is less frequently fixated, and NN and NSUBJ, where the next word is more frequently fixated. Whether the next word is fixated or not, can be influenced by the word length, as well as the fixation probability of the current word: If the word is very short, the next word can be processed from a fixation of the current word, and if the current word is not fixated, the eyes need to land somewhere in order for the visual span to cover a satisfactory part of the text. Word length and fixation probabilities for the nouns are reported in Figure 3c and Figure 3b to show that the dependency labels have similar densities.

Dependency parsing We also evaluate our gaze features directly in a supervised dependency parser. Our baseline performance is relatively low because of the small training set, but comparable to performance often seen with low-resource languages. Evaluation metrics are labeled attachment scores (LAS) and unlabeled attachment scores (UAS), i.e. the number of words that get assigned the correct syntactic head w/o the correct dependency label.

Gaze features lead to consistent improvements across all five domains. The average error reduction in LAS is 5.0%, while the average error reduc-



(a) Next word fixation probability (b) Fixation probability (c) Word length

Figure 3: Kernel density plots across four grammatical functions of nouns.

	LAS						UAS					
					+FREQ+LEN						+FREQ+LEN	
	BL	+SENNA	+EIGENW	+GAZE	BL	+GAZE	BL	+SENNA	+EIGENW	+GAZE	BL	+GAZE
HDL	0.539	0.539	0.526	**0.541**	0.535	**0.542**	0.583	**0.600**	0.564	0.589	0.582	**0.587**
MAI	0.667	0.651	0.668	***0.684**	0.678	***0.711**	0.715	0.699	0.715	***0.747**	0.732	***0.759**
TWI	0.532	0.569	**0.563**	*0.561	0.554	***0.569**	0.576	**0.626**	0.615	*0.602	0.607	***0.621**
WBL	0.604	0.629	0.592	***0.638**	0.631	***0.655**	0.668	0.670	0.666	***0.711**	0.709	***0.719**
WSJ	0.635	0.635	0.622	***0.650**	0.629	**0.634**	0.672	0.681	0.674	***0.695**	0.671	**0.677**
Average	0.595	0.605	0.594	***0.615**	0.605	***0.622**	0.643	0.655	0.647	***0.669**	0.660	***0.672**

Table 3: Dependency parsing results on all five test sets using 200 sentences (four domains) for training and 50 sentences (one domain) for evaluation. Best results are bold-faced, and significant ($p < 0.01$) improvements are asterisked.

tion in UAS is 7.3%. For the +FREQ+LEN experiment, +GAZE also lead to improvements for all domains, with error reductions of 3.3% for LAS and 4.7% for UAS.

For comparison we also ran our parser with SENNA embeddings[5] and EIGENWORDS embeddings.[6] The gaze vectors proved overall more informative.

4 Related work

In addition to Barrett and Søgaard (2015), our work relates to Matthies and Søgaard (2013), who study the robustness of a fixation prediction model across readers, not domains, but our work also relates in spirit to research on using weak supervision in NLP, e.g., work on using HTML markup to improve dependency parsers (Spitkovsky, 2013) or using click-through data to improve POS taggers (Ganchev et al., 2012).

There have been few studies correlating reading behavior and general dependency syntax in the literature. Demberg and Keller (2008), having parsed the Dundee corpus using MINIPAR, show that dependency integration cost, roughly the distance between a word and its head, is predictive of reading times for nouns. Our finding could be a side-effect of this, since NSUBJ, NN and DOBJ/POBJ typically have very different dependency integration costs, while DOBJ and POBJ have about the same. Their study thus seems to support our finding that gaze features can be used to discriminate between the grammatical functions of nouns. Most other work of this kind focus on specific phenomena, e.g., Traxler et al. (2002), who show that subjects find it harder to process object relative clauses than subject relative clauses. This paper is related to such work, but our interest is a broader model of syntactic influences on reading patterns.

5 Conclusions

We have shown that gaze features can be used to discriminate between a subset of grammatical functions, even across domains, using only a small dataset and explored which features are more useful. Furthermore, we have shown that gaze features can be used to improve a state-of-the-art dependency parsing model, even when trained on small seeds of data, which suggests that parsers can benefit from data from human processing.

[5] `http://ronan.collobert.com/senna/`

[6] `http://www.cis.upenn.edu/~ungar/eigenwords/`

Appendix: Gaze features

First fixation duration on every word, fixation probability, mean fixation duration per sentence, mean fixation duration per word, next fixation duration, next word fixation probability, probability to get 1st fixation, probability to get 2nd fixation, previous fixation duration, previous word fixation probability, re-read probability, reading time per sentence normalized by word count, share of fixated words per sentence, time percentage spent on this word out of total sentence reading time, total fixation duration per word, total regression from word duration, total duration of regressions to word, n fixations on word, n fixations per sent normalized by token count, n long regressions from word, n long regressions per sentence normalized by token count, n long regressions to word, n refixations on word, n re-fixations per sentence normalized by token count, n regressions from word, n regressions per sentence normalized by token count, n regressions to word.

References

Maria Barrett and Anders Søgaard. 2015. Reading behavior predicts syntactic categories. In *CoNLL 2015*, pages 345–249.

Michael Collins. 2002. Discriminative training methods for Hidden Markov Models. In *EMNLP*.

Vera Demberg and Frank Keller. 2008. Data from eyetracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109:193–210.

Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Josef Le Roux, Joakim Nivre, Deirde Hogan, and Josef van Genabith. 2011. From news to comments: Resources and benchmarks for parsing the language of Web 2.0. In *IJCNLP*.

Kuzman Ganchev, Keith Hall, Ryan McDonald, and Slav Petrov. 2012. Using search-logs to improve query tagging. In *ACL*.

Franz Matthies and Anders Søgaard. 2013. With blinkers on: Robust prediction of eye movements across readers. In *EMNLP*, Seattle, Washington, USA.

Nicolai Meinshausen and Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Matthias Nilsson and Joakim Nivre. 2009. Learning where to look: Modeling eye movements in reading. In *CoNLL*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Keith Rayner and Susan A. Duffy. 1988. On-line comprehension processes and eye movements in reading. In *Reading research: Advances in theory and practice*, pages 13–66, New York. Academic Press.

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

Jarkko Salojärvi, Ilpo Kojo, Jaana Simola, and Samuel Kaski. 2003. Can relevance be inferred from eye movements in information retrieval. In *Proceedings of WSOM*, volume 3, pages 261–266.

Javier San Agustin, Henrik Skovsgaard, Emilie Mollenbach, Maria Barret, Martin Tall, Dan Witzner Hansen, and John Paulin Hansen. 2010. Evaluation of a low-cost open-source gaze tracker. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 77–80. ACM.

Valentin Ilyich Spitkovsky. 2013. *Grammar Induction and Parsing with Dependency-and-Boundary Models*. Ph.D. thesis, STANFORD UNIVERSITY.

Matthew Traxler, Robin Morris, and Rachel Seely. 2002. Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, 47:69–90.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *ACL*.

Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 188–193. Association for Computational Linguistics.