With blinkers on: robust prediction of eye movements across readers

Franz Matties and Anders Søgaard
University of Copenhagen
Njalsgade 142
DK-2300 Copenhagen S
Email: soegaard@hum.ku.dk

Abstract

Nilsson and Nivre (2009) introduced a tree-based model of persons' eye movements in reading. The individual variation between readers reportedly made application across readers impossible. While a tree-based model seems plausible for eye movements, we show that competitive results can be obtained with a linear CRF model. Increasing the inductive bias also makes learning across readers possible. In fact we observe next-to-no performance drop when evaluating models trained on gaze records of multiple readers on new readers.

1 Introduction

When we read a text, our gaze does not move smoothly and continuously along its lines. Rather, our eyes fixate at a word, then skip a few words, to jump to a new fixation point. Such rapid eye movements are called *saccades*. Sometimes we even jump backwards. Backward saccades are called *regressions*. Gaze can be recorded using eye tracking devices (Starr and Rayner, 2001). Since eye movements in reading give us important information about what readers find complicated in a text, and what readers find completely predictable, predicting eye movements on new texts has many practical applications in text-to-text generation and human computer interaction, for example.

The problem of predicting eye movements in reading is, for a reader r_i and a given sequence of word tokens $w_1 \ldots w_n$, to predict a set of fixation points $F \subseteq \{w_1, \ldots, w_n\}$, i.e., the fixation points of r_i's gaze. For each token w_j, the reader r_i may skip w_j or fixate at w_j. Models are evaluated on recordings of human reading obtained using eye tracking devices. The supervised prediction problem that we consider in this paper, also uses eye tracking data for learning models of eye movement.

Nilsson and Nivre (2009) first introduced this supervised learning task and used the Dundee corpus to train and evaluate a tree-based model, essentially treating the problem of predicting eye movements in reading as transition-based dependency parsing.

We follow Hara et al. (2012) in modeling only forward saccades and *not* regressions and refixations. While Nilsson and Nivre (2009) try to model a subset of regressions and refixations, they do *not* evaluate this part of their model focusing only on fixation accuracy and distribution accuracy, i.e., they evaluate how well they predict *a set of* fixation points rather than a sequence of points in order. This enables us to model eye movements in reading as a sequential problem of determining the length of forward saccades, increasing the inductive bias of our learning algorithm in a motivated way. Note that because we work with visual input, we do not tokenize our input in our experiments, i.e., punctuation does not count as input tokens.

Example Figure 1 presents an example sentence and gaze records from the Dundee corpus. The Dundee corpus contains gaze records of 10 readers in total. Note that there is little consensus on what words are skipped. 5/10 readers skip the first word. Generally, closed class items (prepositions, copulae, quantifiers) seem to be skipped more open, but we do see a lot of individual variation. While others for this reason have refrained from evaluation across readers (Nilsson and Nivre, 2009; Hara et al., 2012),

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 803–807,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

	Sentence									
	Are	tourists	enticed	by	these	attractions	threathening	their	very	existence?
r_1	Fixate	Fixate	Fixate	Skip	Fixate	Fixate	Fixate	Skip	Fixate	Fixate
r_2	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate
r_3	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Skip	Fixate
r_4	Skip	Fixate	Fixate	Skip	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate
r_5	Skip	Fixate	Fixate	Skip	Fixate	Fixate	Fixate	Skip	Fixate	Fixate
r_6	Skip	Fixate	Fixate	Skip	Fixate	Fixate	Fixate	Fixate	Skip	Fixate
r_7	Skip	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate
r_8	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate
r_9	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate
r_{10}	Skip	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Skip	Fixate
# skips	5	0	0	4	0	0	0	2	3	0

Figure 1: The gaze records of the three first readers for the first sentence in the Dundee corpus.

we show that our model predicts gaze better *across readers* than a previously proposed model (Nilsson and Nivre, 2009) does training and evaluating on the *same* readers. A final observation is that fixations are very frequent at the word level – in fact, even skilled readers make 94 fixations per 100 words (Starr and Rayner, 2001) – which motivates using F_1-score of skips as metric. We follow Nilsson and Nivre (2009) in reporting word-level accuracy, but find it particularly interesting that the simple model proposed here outperforms previous models by a large margin in F_1-score over skips.

Related work Below we use a sequential model rather than a tree-based model to bias our model toward predicting forward saccades. Nilsson and Nivre (2009), in contrast, present a more expressive tree-based model for modeling eye movements, with some constraints on the search space. The transition-based model uses consecutive classification rather than structured prediction. The features used in their model are very simple. In particular, they use use word lengths and frequencies, like us, as well as distances between tokens (important in a transition-based model), and, finally, the history of previous decisions.

Hara et al. (2012) use a linear CRF model for the same problem, like us, but they consider a slightly different problem, namely that of predicting eye movement when reading text on a specific screen. They therefore use screen position as a feature. In addition, they use word forms, POS, various measures of surprise of word length, as well as perplexity of bi- and trigrams. The features relating to screen position were the most predictive ones.

2 Our approach

We use linear CRFs to model eye movements in reading. We follow Hara et al. (2012) in using small window sizes (at most five words) for extracting features. Rather than using word forms, POS, etc., we use only word length and the log probability of words – both known to correlate well with likelihood of fixation, as well as fixation times (McDonald and Shillcock, 2012; Kliegl et al., 2004; Reingold et al., 2012). The model thus reflects a hypothesis that eye movements are largely unaffected by semantic content, that eye movements depend on the physical properties and frequency of words, and that there is a sequential dependence between fixation times. Tabel 1 gives the complete set of features. We also evaluated using word forms and POS on held-out data, but this did not lead to improvements. There is evidence for the impact of morphology on eye movements (Liversedge and Blythe, 2007; Bertram, 2011), but we did not incorporate this into our model. Finally, we did not incorporate predictability of tokens, although this is also known to correlate with fixation times (Kliegl et al., 2004). Hara et al. (2012) use perplexity features to capture this.

We use a publicly available implementation of linear CRFs[1] with default parameters (L_2-regularized, $C = 1$).

[1] https://code.google.com/p/crfpp/

3 Predicting a reader's eye movements

In this experiment we consider exactly the same setup as Nilsson and Nivre (2009) considered. In the Dundee corpus, we have gaze data for 10 persons. The corpus consists of 2,379 sentences, 56,212 tokens and 9,776 types. The gaza data was recorded using a Dr. Bouis Oculometer Eyetracker, sampling the position of the right eye every millisecond. We use texts 1–16 (1911 sentences) for training, 17–18 (237 sentences) for development and 19–20 (231 sentences) for testing.

Results are presented in Table 2 and are slightly better than Nilsson and Nivre (2009), mainly because of better predictions of skips. Our error reduction over their model in terms of F_1 over skips is 9.4%. The baseline model used in Nilsson and Nivre (2009), the E-Z Reader (Reichle et al., 1998), obtained a fixation accuracy of 57.7%.

4 Predicting across readers

Hara et al. (2012) consider the problem of learning from the concatenation of the gaze data from the 10 persons in the Dundee corpus, but they also evaluate on data from these persons. In our second experiment, we consider the more difficult problem of learning from one person's gaze data, but evaluating on gaze data from another test person. This is a more realistic scenario if we want to use our model to predict eye movements in reading on anyone but our test persons. This has been argued to be impossible in previous work (Nilsson and Nivre, 2009; Hara et al., 2012).

Our results are presented in Table 3. Interestingly, results are very robust across reader pairs. In fact, only in 4/10 cases do we get the best results training on gaze data from the reader we evaluate on. Note also that the readers seem to form two groups – (a, b, h, i, j) and (c, d, e, f, g) – that provide good training material for each other. Training on concatenated data from all members in each group may be beneficial.

5 Learning from multiple readers

In our final experiment, we learn from the gaze records of nine readers and evaluate on the tenth. This is a realistic evaluation of our ability to predict fixations for new, previously unobserved readers. Interestingly we can predict the fixations of new readers better than Nilsson and Nivre (2009) predict fixations when the training and test data are produced by the same reader. The results are presented in Table 4. In fact our skip F_1 score is actually better than in our first experiments. As already mentioned, this result can probably be improved by using a subset of readers or by weighting training examples, e.g., by importance weighting (Shimodaira, 2000). For now, this is left for future work.

6 Discussion

Our contributions in this paper are: (i) a model for predicting a reader's eye movements that is competitive to state-of-the-art, but simpler, with a smaller search space than Nilsson and Nivre (2009) and a smaller feature model than Hara et al. (2012), (ii) showing that the simpler model is robust enough to model eye movements across readers, and finally, (iii) showing that even better models can be obtained training on records from multiple readers.

It is interesting that a model without lexical information is more robust across readers. This suggests that deep processing has little impact on eye movements. See Starr and Rayner (2001) for discussion. The features used in this study are well-motivated and account as well for the phenomena as previously proposed models. It would be interesting to incorporate morphological features and perplexity-based features, but we leave this for future work.

7 Conclusion

This study is, to the best of our knowledge, the first to consider the problem of learning to predict eye movements in reading across readers. We present a very simple model of eye movements in reading that performs a little better than Nilsson and Nivre (2009) in terms of fixation accuracy, evaluated on one reader at a time, but predicts skips significantly better. The true merit of the approach, however, is its ability to predict eye movements across readers. In fact, it predicts the eye movements of new readers better than Nilsson and Nivre (2009) do when the training and test data are produced by the same reader.

References

Raymond Bertram. 2011. Eye movements and morphological processing in reading. *The Mental Lexicon*, 6:83–109.

Tadayoshi Hara, Daichi Mochihashi, Yoshinobu Kano, and Akiko Aizawa. 2012. Predicting word fixation in text with a CRF model for capturing general reading strategies among readers. In *Workshop on Eye-tracking and NLP, COLING*.

Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16:262–284.

Simon Liversedge and Hazel Blythe. 2007. Lexical and sublexical influences on eye movements during reading. *Language and Linguistic Compass*, 1:17–31.

Scott McDonald and Richard Shillcock. 2012. Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, 14:648–652.

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

Erik Reichle, Alexander Pollatsek, Donald Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological Review*, 105:125–157.

Eyal Reingold, Erik Reichle, Mackenzie Glaholt, and Heather Sheridan. 2012. Direct lexical control of eye movements in reading. *Cognitive Psychology*, 65:177–206.

Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244.

Matthew Starr and Keith Rayner. 2001. Eye movements during reading: some current controversies. *Trends in Cognitive Science*, 5:156–163.

Feature		Description
WordLength	$\{L_{-2}, L_{-1}, L_0, L_1, L_2\}$	The number of letters for a token
WordProbability	$\{P_{-1}, P_0, P_1\}$	The log probability of a word (rounded) as given in the Dundee data

Table 1: Feature template

	Fixation Accuracy		Fixations (F1)		Skips (F1)	
Reader	N&N	Model	N&N	Model	N&N	Model
a	70.0	70.2	71.8	70.0	67.4	70.3
b	66.5	66.2	74.1	71.2	75.0	58.8
c	70.9	70.4	77.3	74.7	59.4	64.4
d	78.9	76.5	84.7	81.3	65.9	68.5
e	71.8	70.5	73.5	69.9	69.9	71.0
f	67.9	66.4	76.8	72.8	47.7	55.8
g	56.6	65.1	61.7	61.8	49.9	67.8
h	66.9	67.7	72.7	70.3	58.2	64.6
i	69.1	71.5	74.1	73.9	60.7	68.8
j	76.3	74.6	82.0	77.3	65.2	71.1
average	69.5	**69.9**	**75.2**	72.3	62.6	**66.1**

Table 2: Comparison between NN09 and our model.

train/test	a	b	c	d	e	f	g	h	i	j
a	-	67.2	67.6	71.5	69.7	63.4	64.9	66.9	70.7	72.6
b	67.7	-	70.1	**76.9**	68.0	65.7	62.9	67.1	69.1	72.8
c	69.3	67.3	-	**76.5**	69.7	65.1	64.3	67.4	71.0	74.2
d	69.0	67.2	70.0	-	69.1	65.1	63.9	67.3	70.1	73.9
e	70.1	66.6	67.5	71.2	-	63.8	64.7	66.9	70.9	72.6
f	66.5	65.9	69.1	**76.7**	66.5	-	62.4	66.8	68.6	71.4
g	69.7	67.1	67.2	69.5	69.6	61.6	-	67.8	70.3	70.3
h	**70.5**	67.5	69.3	74.7	70.5	64.2	64.5	-	70.8	74.2
i	**70.9**	**68.1**	69.6	74.4	**70.7**	64.0	64.6	68.0	-	74.2
j	**70.7**	**68.0**	69.5	74.7	70.4	64.1	64.7	**68.2**	**71.5**	-

Table 3: Results learning across readers. Bold-faced numbers better than when training on same reader

	Fixation Accuracy		Fixations (F1)		Skips (F1)	
Reader	N&N	Model	N&N	Model	N&N	Model
a	70.0	70.3	71.8	72.1	67.4	68.2
b	66.5	67.9	74.1	70.6	75.0	64.6
c	70.9	69.8	77.3	73.1	59.4	65.6
d	78.9	75.5	84.7	79.5	65.9	69.5
e	71.8	70.6	73.5	72.0	69.9	69.0
f	67.9	64.5	76.8	68.6	47.7	59.2
g	56.6	64.7	61.7	65.0	49.9	64.5
h	66.9	68.1	72.7	70.9	58.2	64.8
i	69.1	71.3	74.6	74.1	60.7	67.9
j	76.3	74.2	82.0	77.2	65.2	70.4
average	69.5	**69.7**	**75.2**	72.3	62.6	**66.4**

Table 4: Comparison of NN09 and our cross-reader model trained on nine readers