Using gaze to predict text readability

Ana V. González-Garduño
University of Copenhagen
Department of Computer Science
fcm220@alumni.ku.dk

Anders Søgaard
University of Copenhagen
Department of Computer Science
soegaard@di.ku.dk

Abstract

We show that text readability prediction improves significantly from hard parameter sharing with models predicting first pass duration, total fixation duration and regression duration. Specifically, we induce multi-task Multilayer Perceptrons and Logistic Regression models over sentence representations that capture various aggregate statistics, from two different text readability corpora for English, as well as the Dundee eye-tracking corpus. Our approach leads to significant improvements over Single task learning and over previous systems. In addition, our improvements are consistent across train sample sizes, making our approach especially applicable to small datasets.

1 Introduction

When we read, our eyes move rapidly back and forth between fixations. These movements are called saccades. The distribution of fixations and saccades can provide us with important insight about the reader and the text being read. For example, long regressive eye movements, which typically involve regressing more than 10 letter spaces (Rayner, 1998), may indicate that the reader is facing some difficulty in understanding the text (Frazier and Rayner, 1982; Rayner, 2012). In addition, regressions have been shown to occur during the disambiguation of a sentence (Frazier and Rayner, 1982). This relationship between text and eye movements, has led to an influx of studies investigating the use of eye tracking data to improve and test computational models of language i.e. Barrett et al. (2016); Demberg and Keller (2008); Klerke et al. (2015). In this study we aim to incorporate eye movement data for the task of automatic readability assessment. *Automatic readability assessment* is the task of automatically labeling a text with a certain difficulty level. An accurate and robust system has many potential applications, for example it can help educators obtain appropriate reading materials for students with normal learning capacities, as well as students with disabilities and language learners. It can also be used to assess the performance of machine translation, text simplification and language generation systems. Eye-tracking data has previously been used to evaluate readability models (Green, 2014; Klerke et al., 2015), however, our main contribution is to explore the way that eye tracking data can help improve models for readability assessment through multi-task learning (Caruana, 1997) and parser metrics based on the surprisal theory of syntactic complexity (Hale, 2001, 2016). Multi task learning allows the model to learn various tasks in parallel and improve performance by sharing parameters in the hidden layers.

The work most related to ours is by Singh et al. (2016), who used eye tracking measures taken from the Dundee corpus in order to predict word by word reading times for each sentence. Subsequently, they used these word by word reading times as features for predicting readability. The two tasks were performed separately, and their feature representations were different from the ones presented here. In contrast, we present a model that predicts gaze *and* sentence-level readability simultaneously.

We use gaze data from the Dundee corpus (Kennedy et al., 2003) and two different datasets for the readability prediction task: aligned Wikipedia sentences used for the task of text simplification by Coster and Kauchak (2011) and the OneStopEnglish dataset used by Vajjala and Meurers (2014).

Contributions This is, to the best of our knowledge, the first application of multi-task learning to readability prediction. Our model is also different from previous applications of multi-task learning to natural language processing in that we combine a classification task and a regression task. We experiment with two multi-task learning algorithms, namely hard parameter sharing for multi-layered perceptrons (Caruana, 1997) and a novel approach to hard parameter sharing between logistic and linear regression. We evaluate our models on Simple Wikipedia and the OneStopEnglish corpus. Finally, we present learning curves that show that the improvements are robust across different sample sizes.

2 Experiments

Data Our target task is sentence-level readability prediction, i.e. a binary classification problem of sentences into easy-to-read and hard-to-read.

Our main corpus is a sentence-aligned corpus of 137,000 simple versus normal English sentences from Wikipedia (Coster and Kauchak, 2011). Similar datasets have been used in the past, e.g., in Ambati et al. (2016) and Hwang et al. (2015). The easy-to-read sentences were taken from Simple Wikipedia and paired with sentences from the standard Wikipedia using cosine similarity.

In addition, we also evaluate our models on the OneStopEnglish corpus (Vajjala and Meurers, 2014), specifically the elementary-intermediate and elementary-advanced sentence pairs. This dataset has been used for readability assessment (Vajjala and Meurers, 2014) using the WeeBit model presented by (Vajjala and Meurers, 2012), so we compare our results with theirs.

Feature representation In this study, features known to affect the complexity of text, such as syntactic, lexical and total surprisal (Hale, 2001; Demberg and Keller, 2008), were used. Most of these features were extracted using a probabilistic top-down parser introduced by Roark (2001). After removing duplicate sentences and sentences with typos, the final corpus used was of about 80,000 sentence pairs. The features extracted are shown in table 1.

The *prefix probability* of word w_n is explained by Jelinek and Lafferty (1991) as the probability that w_n occurs as a prefix of some string generated by a grammar. It is the sum of the probabilities of all trees from the first word to the current word. *Surprisal* is then the difference between the log of the prefix probability of w_n and w_{n-1}.

Features
1. Prefix probability -word1
2. Total surprisal - word1
3. Syntactic Surprisal -word1
4. Lexical Surprisal - word1
5. Ambiguity - word1
6. Prefix probability -word2
7. Total surprisal - word2
8. Syntactic Surprisal - word2
9. Lexical Surprisal - word2
10. Ambiguity - word2
11. Total surprisal – sent mean
12. Syntactic Surprisal – sent mean
13. Lexical Surprisal – sent mean
14. Ambiguity – sent mean
15. Total surprisal – sent sd
16. Syntactic Surprisal – sent sd
17. Lexical Surprisal – sent sd
18. Ambiguity – sent sd
19. Sentence length
20. Ave. Word length
21. Parse Tree height
22. Num of Subordinate clauses(SBARs)
23. Num of Noun phrases
24. Num of Verb phrases
25. Num of Prepositional phrases
26. Num of Adv phrases
27. Ratio nouns
28. Ratio verbs
29. Ratio adjectives
30. Ratio pronouns
31. Ratio adverbs
32. Ratio Det
33. Mean Age of Acquisition

Table 1: Features extracted for the readability data.

If we describe $\mathcal{D}(G, W[1,n])$ as the set of all possible leftmost derivations D with respect to probabilistic context free grammar G and whose last step used a production with terminal W_n. We can then express the prefix probability of $W[1,n]$ with respect to G as $PP_G(W[1,n]) = \sum_{D \in \mathcal{D}(G,W[1,n])} \rho(D)$, where $\rho(D)$ is the probability of the derivation of a certain tree.

The total surprisal of W_n is then defined as:

$$S_G(W_n) = -\log \frac{PP_G(W[1,n])}{PP_G(W[1,n-1])}$$

Syntactic surprisal and *lexical surprisal* are calculated to account for high surprisal scores (Roark et al., 2009). As Roark et al. (2009) mentions, a word may surprise because it is unconventional, or because it occurs in an unusual context.

In order to separate the lexical and syntactic components of surprisal, the incremental parser calculates partial derivations immediately before word W_n is integrated into the syntactic structure. Syntactic surprisal ($SynS_G(W_n)$) is defined as:

$$-\log \frac{\sum_{D \in \mathcal{D}(G,W[1,n])} \rho(D[1,|D|-1])}{PP_G(W[1,n-1])}$$

and lexical surprisal ($LexS_G(W_n)$) as:

$$-\log \frac{PP_G(W[1,n])}{\sum_{D \in \mathcal{D}(G,W[1,n])} \rho(D[1,|D|-1])}$$

Where $D[1,|D|-1]$ is the set of the partial derivations before each word is integrated into the structure $\mathcal{D}(G,W[1,n])$. Total surprisal turns out to be sum of syntactic surprisal and lexical surprisal.

We also obtain an entropy score using the parser. Entropy over a set of derivations $\mathcal{D}$, denoted as $H(\mathcal{D})$, quantifies the uncertainty over the partial derivations. We call this feature *Ambiguity*, defined as:

$$-\sum_{D \in \mathcal{D}} \frac{\rho(D)}{\sum_{D' \in \mathcal{D}} \rho(D')} \log \frac{\rho(D)}{\sum_{D' \in \mathcal{D}} \rho(D')}$$

Furthermore, features corresponding to the first and second words were included, as the initial words in a sentence allow the reader to make preliminary guesses of what the structure will be for the rest of the sentence, although these predictions can often turn out to be wrong. In addition, mean syntactic scores and standard deviations for all words in the sentence are included as features. We also include the mean age of acquisition for the words in a given sentence, using data from Kuperman et al. (2012). Finally, we include basic counts and ratios used previously in readability prediction such as sentence length, parse tree height, number of SBAR's, noun phrases, verb phrases, among others .

In order to predict gaze, we extract the features seen in Table 1 from the Dundee corpus. As mentioned earlier, these features offer a good representation of cognitive load, which is also reflected in reading times. A feature vector of size 33 was built for each sentence, and this information was used in order to predict an average first pass duration, regression path duration and total fixation duration.

First pass duration refers to the sum of all fixations on a region once the region is first entered until it is left. *Regression path duration* includes regressions made out of a region prior to moving forward in the text and *total fixation duration* is the sum of all fixations in the region including, regressions to that region. As mentioned in Rayner et al. (2006), these measures typically concern research questions focusing on sentence or discourse processing.

Logistic/linear regression and MLPs Logistic Regression (LR) models have been widely used in document level readability classification i.e. Feng et al. (2010) and (Xia et al., 2016). LR models are linear models and can be thought of as single-layer perceptrons with softmax or sigmoid activation functions. The objective is typically to minimize a cross-entropy loss function. The same architecture can be used for linear regression, however, when trained to minimize mean squared error. Here, we compare LR with a 3-layered Multi Layer Perceptron (MLP). For our MLP architecture, we use sigmoid activation at the input and output layers and use ReLu activation in the hidden layer. The hidden layer contains 100 neurons. All models presented here use the Adam optimizer, and a drop-out rate of 0.5. We also use Adam to learn logistic and linear regression models.

As already mentioned, we go beyond single-task LR and MLP models and present two multi-task learning architectures with heterogeneous loss functions (cross-entropy and minimum squared error). In multi-task learning (Caruana, 1997), the training signals of one task are used as an inductive bias in order to improve the generalization of another task. Specifically, we use the the task of gaze prediction in order to improve the generalization of readability prediction.

Multi-task MLP Our multi-task learning architecture is identical to that of Caruana (1997) and Collobert et al. (2011), i.e., two MLPs that share all parameters in their hidden layers. The only difference is that one of the MLPs in our case is trained to minimize a minimum squared error to predict gaze statistics.

Multi-task logistic and linear regression Our linear multi-task learning model is novel in that it combines a logistic and a linear regression model by tying their parameters. As mentioned earlier, LR models can be thought of as single-layer perceptrons. We tie a single-layer perceptron with sigmoid activation to another single-layer perceptron with linear activation by sharing their single layer and giving a higher weight to our main task. While this is in fact a simpler model than the deep multi-task learning model above, this model has, to the best of our knowledge, not been suggested before, and in many ways, it is surprising that it works.

Baselines Ambati et al. (2016) obtained 78.87 percent accuracy on the Wikipedia dataset. They use features extracted from a Combinatory Categorical Grammar (CCG) parser. We also compare our results to Vajjala and Meurers (2014), who use their WeeBit model in order to predict readability at the sentence level. In addition, for the Wikipedia dataset, we include the best results from Singh et al. (2016) as it is the study most related to ours.

3 Results

Our results are shown in Table 2.

For the Wikipedia corpus, our best multi-task learning system shows an improvement in accuracy over previous work by about 8%. A big part of the improvement can be attributed to using a deep learning architecture. Single-task MLPs do about 8% better than logistic regression on this dataset, in absolute numbers. Multi-task learning buys us another .5%, absolute. For the advanced-elementary sentence pairs in the OneStopEnglish corpus, a slightly larger improvement is seen from multi-task learning to single-task. For all multi-task systems, there is an improvement over the corresponding single-task system with at least two of the gaze inputs. The best result was achieved using Multi-Task MLP. Inclusion of gaze data improved our results about 2.6 % over the best single-task result.

We performed various paired T tests in order to assess whether or not the improvements obtained using multi-task learning was significant. We compared the results of each MTL model to its corresponding STL model. We report p values using asterisks in Table 2. P values smaller than 0.001 are described using ***, ** indicate p values ranging from 0.001 to 0.01 and p values from 0.01 to 0.5 are shown with *. No asterisk indicates that there was no statistically significant changes.

Systems	Wikipedia	OSE (A-E)	OSE (I-E)
Previous			
Singh	75.21	-	-
Ambati	78.87	-	-
Vajjala	66.00	61.0	51.0
Single-task			
LR	78.17	67.23	58.72
MLP	85.95	67.53	59.30
Multi-task LR			
1st pass	78.20	67.88	60.71***
Regression	78.15	68.10*	59.68
Total fix	78.60*	68.08**	60.80**
Multi-task MLP			
1st pass	86.13	68.08 ***	61.70 ***
Regression	86.11	67.66	**61.91***
Total fix	**86.45***	**68.51** ***	61.27 ***

Table 2: Accuracy for all models. Most results obtained using MTL-MLP yield statistically significant improvements of STL-MLP ($p < 0.001$).

4 Discussion

Performance of features The features extracted using the probabilistic top down parser have previously been used in order to predict word by word reading times (Singh et al., 2016; Demberg and Keller, 2008), but have not been thoroughly explored in the task of readability prediction. Here, we used surprisal and entropy, along with other low-level features in order to predict the reading level of single sentences. Using the STL-MLP, we predicted readability using feature groups, separated by syntactic features and low level features. Our syntactic features include features 1-18, while our low level features are 19-33 in table 1. We found that low level features are more predictive for our datasets than syntactic features, however, it is a combination of both that yields the best results. These results can be seen in table 3.

Feature Set	Wikipedia	OSE I-E	OSE A-E
Syntactic Features	60.35	54.90	59.79
Low level Features	78.15	57.30	65.17
All Features	**85.95**	**59.30**	**67.53**

Table 3: Accuracy when predicting readability using features in groups. The results show, that a combination of both sets of features provide the best result.

In addition, we performed a single feature evaluation, where each feature was used to predict readability using the STL-MLP model. The 10 most predictive features for the Wikipedia dataset are presented in table 4. The results reaffirm the previous finding that although syntactic features are predictive of readability, low level features remain the most predictive.

Wikipedia	
Feature	Accuracy
Ratio Verbs	67.90
Ratio Adjectives	66.46
Sentence Length	61.96
Ratio Adverbs	57.53
Mean Age of Acquisition	57.34
Average Word Length	56.17
Ambiguity – sent SD	55.66
Lexical suprisal – sent SD	55.37
Num of verb phrases	55.13
Ambiguity Sent Mean	55.00

Table 4: Accuracy on Wikipedia dataset when predicting readability using single features.

Effects of using gaze data The main objective of this study was to explore how the use of eye tracking data improved our readability prediction model. Using the Dundee eye tracking corpus, we were able to learn models for predicting an average first pass duration, total regression to duration, and total fixation duration for a given sentence in our readability datasets. Using hard parameter sharing, we learned to predict a readability label and gaze simultaneously. This method allows us to exploit the information contained in one task to better generalize another. Our results demonstrate that gaze data does improve readability models significantly.

Learning curves In figure 1 we compare the learning curves for the best MTL and STL models for each dataset. We show the accuracy on both validation sets using varying amounts of train samples. The first train sample used consisted of 100 sentences. At this small sample size, the effect of the gaze data is more clear. For example, for the Wikipedia dataset the validation accuracy using 100 samples is about 74.5 % for the MTL MLP system, while for the STL MLP system, the accuracy is about 10 % lower. At about 20,000 samples the difference in performance between the two systems begins to level off, however, MTL remains slightly higher the entire time. This is in line with Caruana (1997), who mentions that the improvements using MTL are typically stronger when using smaller sample sizes.

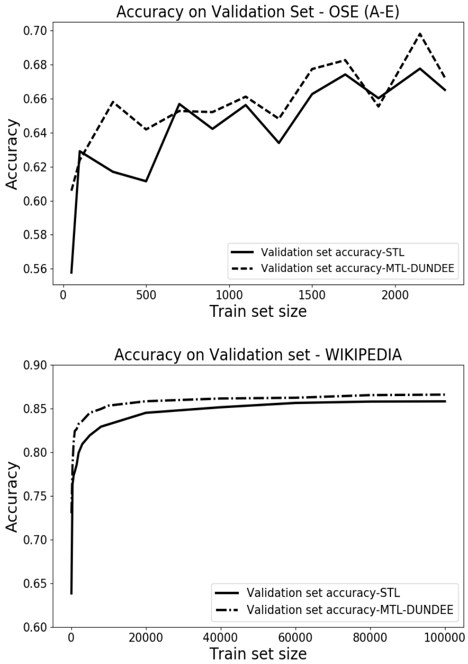

Figure 1: Learning curves for the OSE A-E and Wikipedia datasets varying the train sample size. The first sample size consisted of 100 sentences.

Similar results can be seen for the Advanced-Elementary sentence pairs. We begin training our model on about 100 samples and incrementally increased the train set size. Neither of the models achieve high accuracy, however, the MTL system improves the result about 5 %, and as the training set size increases, this trend persists. Similar results are observed for the Intermediate-Advanced pairs.

5 Conclusion

In this study, we have presented the first application of multi-task learning to predicting sentence-level readability. We presented two models: a deep learning model and a linear model. The linear multi-task learning model is novel and yields statistically significant results, however, the deep learning model performs better. We present a learning curve analysis showing that multi-task learning is more effective with small sample sizes, however, the improvements are robust across sample sizes.

References

Ram Bharat Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental ccg parser. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1057.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(41).

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

William Coster and David Kauchak. 2011. Simple english wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers*, pages 665–669.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):192–210.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics.

Lyn Frazier and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, 14(2):178–210.

Matthew J. Green. 2014. An eye-tracking evaluation of some parser complexity metrics. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, page 38–46. Association for Computational Linguistics.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9).

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from standard wikipedia to simple wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).*

Frederick Jelinek and John D. Lafferty. 1991. Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics*, 17(3):315–323.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The dundee corpus. *Proceedings of the 12th European conference on eye movement.*

Sigrid Klerke, Sheila Castilho, Maria Barrett, and Anders Søgaard. 2015. Reading metrics for estimating task efficiency with mt output. In *Conference on Empirical Methods in Natural Language Processing*, page 6.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4).

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

Keith Rayner. 2012. *Eye movements in reading: Perceptual and language processes*. Elsevier.

Keith Rayner, Kathryn H Chace, Timothy J Slattery, and Jane Ashby. 2006. Eye movements as reflections of comprehension processes in reading. *Scientific studies of reading*, 10(3):241–255.

Brian Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2).

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, page 324–333. Association for Computational Linguistics.

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajakrishnan Rajkumar. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, page 202–212.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *The 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, page 163–173.

Sowmya Vajjala and Detmar Meurers. 2014. Readability assessment for text simplification. *International Journal of Applied Linguistics*, 165(2).

Menglin Xia, Ekaterina Kochmar, and E Briscoe. 2016. Text readability assessment for second language learners.