The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Learning to Predict Readability Using Eye-Movement Data from Natives and Learners

Ana V. González-Garduño, Anders Søgaard
Department of Computer Science, University of Copenhagen
{ana, soegaard}@di.ku.dk

Abstract

Readability assessment can improve the quality of assisting technologies aimed at language learners. Eye-tracking data has been used for both inducing and evaluating general-purpose NLP/AI models, and below we show that unsurprisingly, gaze data from *language learners* can also improve multi-task readability assessment models. This is unsurprising, since the gaze data records the reading difficulties of the learners. Unfortunately, eye-tracking data from language learners is often much harder to obtain than eye-tracking data from native speakers. We therefore compare the performance of deep learning readability models that use *native speaker* eye movement data to models using data from language learners. Somewhat surprisingly, we observe no significant drop in performance when replacing learners with natives, making approaches that rely on native speaker gaze information, more scalable. In other words, our finding is that *language learner difficulties can be efficiently estimated from native speakers*, which suggests that, more generally, readily available gaze data can be used to improve educational NLP/AI models targeted towards language learners.

Introduction

Automatic readability assessment is a common task in NLP/AI and refers to the task of predicting reading difficulties. Some work in automatic readability assessment is directed at predicting reading difficulties of native speakers with normal reading abilities, but there is also a growing literature using readability assessment to select appropriate literature for language learners.

Vajjala and Meurers (2012), for example, found that developmental measures from second language acquisition research, in combination with traditional features such as word length and sentence length, improve document level readability assessment. More recently, Xia, Kochmar, and Briscoe (2016) used self-training to adapt models of text readability trained on native speaker data to estimate readability for learners.

In parallel, recent studies in text readability have made use of eye movement data from native speakers of English in order to evaluate their models (Klerke et al. 2015; Green 2014). These studies attempt to take into account the correlation between eye-tracking measures such as first pass duration and regression duration, and perceived text difficulty (Rayner et al. 2012).

The work most similar to ours is by Singh et al. (2016), who presents a two-level system to predicting the readability of Wikipedia sentences. Singh et al. (2016) used reading times from the Dundee eye-tracking corpus to learn to predict reading times in the first level of the system and then used these as features for predicting readability in the second. They found this to be a successful approach, however, their results were slightly below the state-of-the-art (Ambati, Reddy, and Steedman 2016).

In contrast to Singh et al. (2016), our study focuses on the effects of using eye movement data from native speakers versus learners of English, in a multi-task learning set-up (Caruana 1997), with text readability assessment as our main task. In multi-task Learning, the training signals of one task, the auxiliary task, are used to improve the performance of the main task, by sharing information throughout the training process. Specifically in this study, we use the task of eye-movement prediction to *induce* models of readability. Furthermore, we compare using eye-movement prediction for native speakers as an auxiliary task, to using gaze prediction for language learners. In both cases, we use the auxiliary task to induce a multi-task Multi-Layered Perceptron model for text readability assessment.

We evaluate our models on both Simple Wikipedia vs Normal Wikipedia (Coster and Kauchak 2011) and OneStopEnglish (Vajjala and Meurers 2014) corpus. We use prefix probabilities, surprisal, ambiguity scores as well as other low-level features typically used in readability assessment in order to represent each sentence of the corpora. The gaze data that we use for our auxiliary tasks comes from the Dundee corpus (Kennedy and Pynte 2003), as well as the recently introduced Ghent Eye-Tracking Corpus (GECO) (Cop et al. 2017), which includes both gaze measures from native speakers and learners of English.

Contributions This is, to the best of our knowledge, the first application of multi-task learning to readability prediction, comparing data from natives speakers versus learners of English. We begin with the observation that, unsurprisingly, gaze data of language learners can be used to improve readability assessment for language learners. This is unsurprising, because gaze data records the reading behavior, including the length of fixations and refixations, of language learners, directly reflecting their processing times, as well as when they need to revisit passages to understand the texts they read. Our results show the effects from using gaze data are robust across training sample sizes and across datasets. More surprisingly, however, we also show that similar effects can be obtained using gaze data from *native readers*. This is interesting, since it is much easier to collect data from native readers than from language learners, especially for low-resource languages.

Experiments

Data Our main task is to improve the performance of our text readability predictions. For this purpose, we use two corpora that have been previously used in readability assessment. The first is a sentence aligned corpus of 137,000 simple versus normal English sentences (Coster and Kauchak 2011), which was made in order to assess the performance of simplification systems. The dataset contains sentences from Simple Wikipedia and standard Wikipedia that were paired using cosine similarity.

The second corpus is the sentence-level OneStopEnglish Corpus (Vajjala and Meurers 2014), which consists of sentences at three different levels: Elementary, Intermediate and Advanced. This corpus was annotated by experts who read a news article and simplified it into two levels: intermediate and elementary. We used Elementary-Advanced and Elementary-Intermediate sentence pairs in our experiments.

For the auxiliary task of predicting eye movements, we use the Dundee Corpus (Kennedy and Pynte 2003), which has been used in readability studies and studies of syntactic complexity (Singh et al. 2016; Demberg and Keller 2008). The English portion of the Dundee corpus consists of eye-tracking measures taken from 10 native speakers of English while reading newspaper articles from *The Independent*. It contains a total of 56,212 tokens and 2,368 sentences and a total of 9,776 types.

In addition, we use the GECO corpus (Cop et al. 2017), which consists of data from 14 monolingual native English speakers, and 19 native speakers of Dutch with English as their second language. All subjects read a literary novel. The 19 participants varied in their level of proficiency in English, from lower-intermediate to advanced.

For our study, we use the gaze data collected from the English native speakers, as our L1 data, and the gaze data collected from the Dutch speakers while reading in English, as our L2 data. For each sentence in the eye movement corpora, we extract three eye movement measures that are commonly used when investigating sentence processing (Rayner et al. 2006): first pass duration, regression path duration and total fixation duration.

First pass duration refers to the time spent reading a word the first time the gaze enters the corresponding visual region. *Total regression duration* describes the total time spent in an area after the gaze has left a word's visual region for the first time. Regressive eye movements have shown to be correlated with comprehension difficulty as well as perceived reading difficulty (Rayner et al. 2012). These two measures can be seen as early and late processing measures. In addition, we predict *total fixation duration*, which refers to the sum of all fixation durations. Specifically, in our study, we extract the *average* first pass, regression path and total fixation durations for the words in a given sentence.

For all experiments, we use 60% of the data for training, 20% for development and 20% for testing.

Features We use various features known to affect text complexity, including syntactic, lexical and total surprisal measures, extracted from a probabilistic top-down parser (Roark 2001).

1. Prefix probability -word 1
2. Total surprisal - word 1
3. Syntactic surprisal - word 1
4. Lexical surprisal - word 1
5. Ambiguity - word 1
6. Prefix probability -word 2
7. Total surprisal - word 2
8. Syntactic surprisal - word 2
9. Lexical surprisal - word 2
10. Ambiguity - word 2
11. Total surprisal sentence mean
12. Syntactic surprisal sentence mean
13. Lexical surprisal sentence mean
14. Ambiguity sentence mean
15. Total surprisal sentence SD
16. Syntactic surprisal sentence SD
17. Lexical surprisal sentence SD
18. Ambiguity sentence SD
19. Sentence length
20. Average word length
21. Parse tree height
22. # of Subordinate clauses (SBARs)
23. # of Noun phrases
24. # of Verb phrases
25. # of Prepositional phrases
26. # of Adv phrases
27. Ratio nouns
28. Ratio verbs
29. Ratio adjectives
30. Ratio pronouns
31. Ratio adverbs
32. Ratio determiners
33. Mean age of acquisition

Table 1: Feature set for each sentence.

The *prefix probability* of word w_n ((Jelinek and Lafferty 1991)) is 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* can be derived from the prefix probability by taking the difference between the log of the prefix probability of w_n and w_{n-1}.

If $\mathcal{D}(G, W[1, n])$ is 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 describe 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.

Syntactic surprisal and *lexical surprisal* are calculated to account for high surprisal scores (Roark et al. 2009), due to a word appearing in an unusual context or because it is uncommon. The incremental parser, isolates the syntactic and lexical components of Surprisal by calculating the 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$\mathcal{D}(G,W[1,n])$. The sum of syntactic surprisal and lexical surprisal is the total surprisal. In addition, the parser extracts an entropy score. Entropy over a set of derivations $\mathcal{D}$, denoted as $H(\mathcal{D})$, quantifies the uncertainty over the partial derivations. We call this feature as *Ambiguity*, defined as:

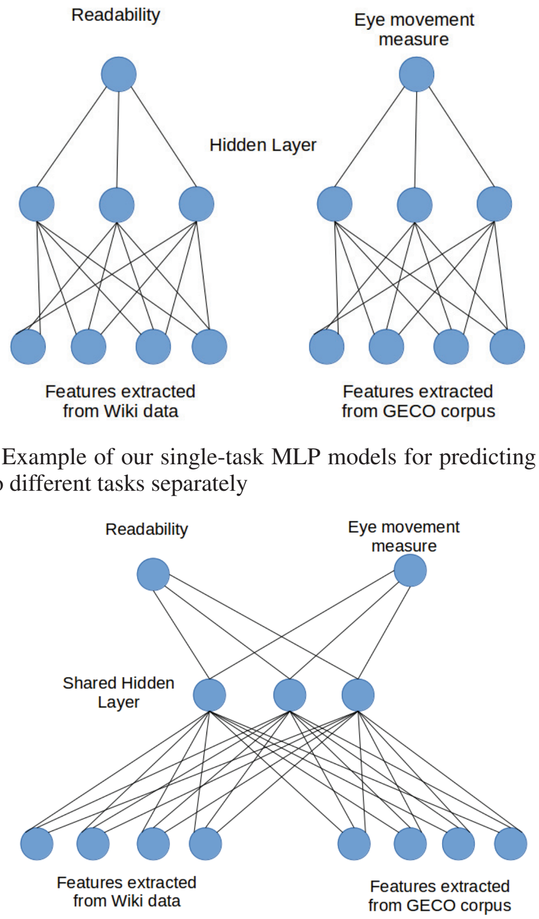
$$-\sum_{D\in\mathcal{D}} \frac{\rho(D)}{\sum_{D'\in\mathcal{D}} \rho(D')} \log \frac{\rho(D)}{\sum_{D'\in\mathcal{D}} \rho(D')}$$

In addition, we include low-level features known to affect readability such as sentence length, average word length, parse tree height and number of subordinate clauses (SBAR). The full set of features is seen in Table 1. We extracted the same set of features for each sentence in the readability and eye-tracking corpora.

MLP and Multi-Task MLP For the single-task system, we use a three-layer perceptron with sigmoid activation at the output layer for readability prediction and linear activation for eye movement prediction models. We use ReLu activation in the hidden layers which contains 100 neurons. All models use Adam optimizer and a drop-out rate of 0.5. In figure 1a, we can observe the single-task MLP architecture for both readability prediction and eye movement prediction. For the task of predicting readability, the model takes as input the features extracted from each sentence in the readability corpora (Wikipedia or OSE) and predicts a binary label corresponding to easy vs difficult sentence. The eye movement model takes in the features extracted from the sentences in the Dundee and GECO corpora and predicts a value corresponding to the average eye movement duration for that sentence.

In addition to the single-task system, we use a multi-task Learning architecture identical to that of (Collobert et al. 2011). This consists of two MLP's with the same architecture as the single-task system, that in addition share all parameters in their hidden layers. One MLP minimizes a mean squared error in order to predict gaze. Figure 1b shows the architecture of out multi-task MLP model. The model takes two inputs corresponding to the features extracted from one readability dataset and one eye movement dataset, and outputs two values, a readability label and one of the eye movement measures. The model is fully connected at the hidden layer, where parameters are shared between the two tasks.

As mentioned earlier, the main task is predicting readability and the auxiliary task is eye movement prediction, therefore, in this study we focus our evaluations on our readability results only. However, because multi-task learning implies that the two tasks are related, we present briefly our single-task MLP results for our eye movement predictions.


Readability
Eye movement measure
Hidden Layer
Features extracted from Wiki data
Features extracted from GECO corpus
Readability
Eye movement measure
Shared Hidden Layer
Features extracted from Wiki data
Features extracted from GECO corpus

(a) Example of our single-task MLP models for predicting the two different tasks separately

(b) Example of our multi-task MLP model which predicts the two different tasks simultaneously

Figure 1

Baselines We include the results from Ambati, Reddy, and Steedman (2016) as a baseline for our Wikipedia results, as well as the results from (Vajjala and Meurers 2014) as a baseline for the OSE sentence pairs. Both of these results were obtained using single-task Learning architectures. In addition, we compare our readability prediction results when using our multi-task MLP to our single-task MLP as our baseline for all datasets.

Results

The results for the task of readability prediction for all systems are shown in Table 2. For all datasets, multi-task systems using the GECO dataset gave the best results. Improvements over single-task baselines were in the range of .67–3.35%. Improvements over using the Dundee corpus were in the range of .17–.73%. Our best results are significantly better than both previous work and our single-task baselines ($p < 0.01$). See the results table for significance of all results relative to the single-task baseline.

As mentioned earlier, multi-task Learning assumes that the tasks are related so that learning one task can actually improve the performance of the main task. In our case, we wanted to know whether the feature representation used to predict both tasks was successful in a single-task set up, therefore, we include the results of our single-task MLP models for predicting eye movements using the extracted features. Our results, shown in Table 3, show the correlation between our single-task MLP predictions versus the true values. All correlations are statistically significant. We do not evaluate our multi-task MLP eye movement predictions in this study.

Systems	Wikipedia	OSE (A-E)	OSE (I-E)
Previous			
Ambati	78.87	-	-
Vajjala	-	61.0	51.0
Single-task MLP			
No Gaze	85.95	67.53	59.30
Multi-task MLP -Dundee			
1st pass	86.13	68.08*	61.70**
Regression	86.11	67.66	61.91**
Total fix	86.45**	68.51**	61.27**
Multi-task MLP -L1 (GECO)			
1st pass	86.51**	68.77**	**62.64****
Regression	86.41*	**69.18****	62.45**
Total fix	86.58*	68.57*	61.43*
Multi-task MLP -L2 (GECO)			
1st pass	**86.62****	68.57*	61.84**
Regression	86.58**	68.37*	62.25**
Total fix	86.35	68.97**	61.63**

Table 2: This table shows the accuracy for all multi-task and single-task systems. In all experiment, 60 % of the data was used for training, 20 % for development and 20 % for testing. Significance is indicated with the asterisks: $** = p < 0.01$, $* = 0.01 \geq p < 0.05$.

Discussion

Effect of using gaze data We compared the learning curves of the best performing multi-task and single-task systems when varying the amount of training data. For the single-task models, began training on 100 sentences and incrementally increased until we were using 60 percent of the data for training. For the MTL models, we started with 100 sentences taken from the readability corpora and 100 sentences takes from the eye movement corpora and incrementally increased until 60 percent of the readability data was used for training. To deal with the difference in corpus size, we copied the eye movement sentences until the size of the corpus matched the size of the readability corpus.

Gaze measure	Pearson's ρ
Dundee	
First pass	0.45***
Regression duration	0.56***
Total fixation duration	0.57***
GECO -L1	
First pass	0.49***
Regression duration	0.59***
Total fixation duration	0.51***
GECO -L2	
First pass	0.65***
Regression duration	0.65***
Total fixation duration	0.62***

Table 3: Pearson's ρ correlation coefficient between the predictions of the MLP models and the true values. All our results have a p-value lower than 0.001 (***).

Subset	Improvements	Losses
L1	14551	3757
L2	16189	4230

Table 4: This table shows the number of improvements and losses when using multi-task learning and single-task learning with the different samples (L1 and L2 speakers)

Figure 2 shows the learning curves for the best performing single-task and multi-task systems. We do not include the multi-task MLP-Dundee system because all other MTL systems outperform it. Across the three sets of sentence pairs, we see that all multi-task systems perform better than the single-task systems. Furthermore, the systems that use L2 gaze data to learn an auxiliary task show to have similar effects when using smaller sample sizes. For example, when the training sample size is as small as 100 samples, both L1 and L2 still generalize much better than the single-task system. For the Wikipedia dataset, both L1 and L2 systems are about 10% more accurate than the single-task system. For the other two data subsets, the difference is not as large, however, a significant difference can still be observed, when training multi-task and single-task systems on 100 samples.

Differences between L1 and L2 As mentioned earlier, training all models on 100 samples from the Wikipedia dataset, the multi-task models perform about 10% better than single-task models. In order to further examine this substantial improvement, we labeled each instance where

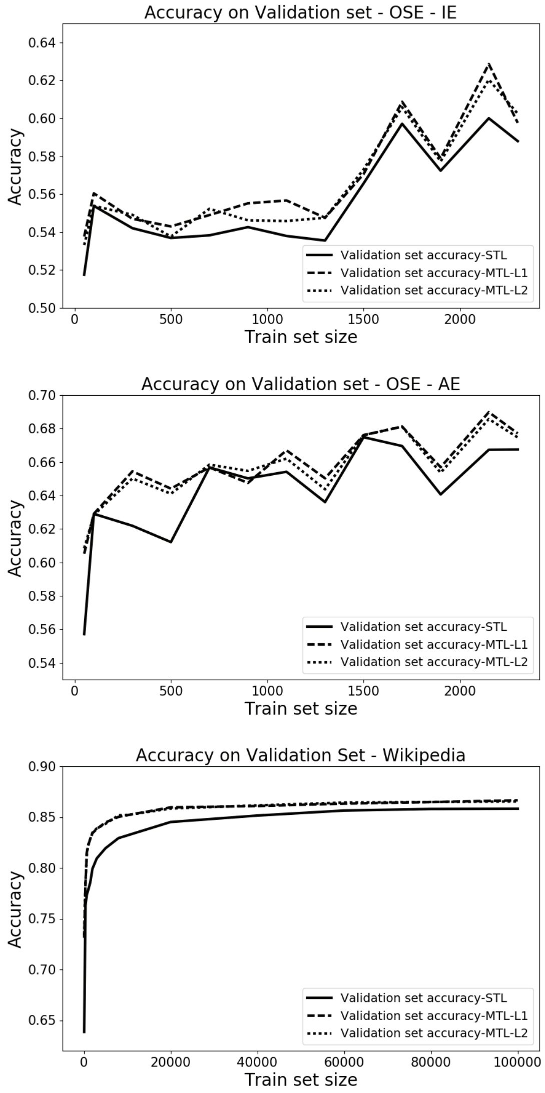


Figure 2: Learning curves using varying amounts of training data

the multi-task learning output improved over the single-task output. Conversely, when our single-task models performed better than the multi-task model, we labeled that with a different label. Instances where there was no difference in performance, were removed. The counts of improvements versus losses when using multi-task learning over single tasks, for the Wikipedia dataset, are seen in Table 4.

Splitting our data the same way as in the previous experiments, we then trained a decision tree classifier on single features to discriminate *between instances where results improve from multi-task learning, and instances where multi-task learning hurts*, for the L1 versus the L2 datasets. We list features that predict multi-task learning gains with an F1-score higher than 75% in Table 5. This suggests that several features are highly predictive of gains. Note that sentence length is also very predictive, and that many of our features covary with sentence length.

Generally, the results show that when using both datasets, low-level features remain the most predictive. For L1 gaze data, we also see parser metrics that are predictive of improvements and losses. Specifically, we observe that syntactic, lexical and total surprisal, as well as ambiguity of first word achieve F1-scores higher than 75 %. In the case of the L2 eye movement data, however, only syntactic surprisal has an F1-score higher than 75%.

This pattern possibly reflects the fact that the eye movement data for the L2 readers contains more variability, i.e., the readers are at different proficiency levels of English, and that many beginner or intermediate L2 readers are more sensitive to low-level features such as sentence length.

Inspecting the sentences that improved from incorporating eye movement prediction as an auxiliary task, we see a lot of improvements concern sentences that are either very short or long. When complexity was a result of subtle vocabulary changes, improvements were rarer. In Table 6, we show examples of the type of sentences that lead to improvements and losses in the best multi-task MLP model.

Eye movement Prediction In multi-task learning it is important that the tasks that are learned simultaneously are related to a certain extent (Caruana 1997; Collobert et al. 2011). As eye movement behavior is known to correlate with text difficulty (Rayner et al. 2012; Clifton, Staub, and Rayner 2007), we know that both our main and auxiliary tasks should in fact be related, however, it is important to establish whether the same feature representation can yield good results for both tasks independently.

We also evaluate the performance of our eye movement predictions. We randomly select 50 samples and visualize the results in Figure 3. The plot shows that for both the L1 and L2 subsets, the single-task MLP model is successfully predicting eye movements, however, the eye movements of L2 readers are slightly more predictable. This is observed both in the plots, as well as the correlation results in Table 3, however, this does not seem to significantly improve the results in the multi-task MLP models that use L2 eye movement information over the models that incorporate L1 eye movement information. Since the L2 readers are reported to have different proficiency levels, it may be surprising that their gaze patterns are more predictable, but we speculate this is balanced by higher intra-reader variance leading to a clearer signal.

Gaze information from natives can be successfully used when L2 data is not readily available, as the benefits might be similar for both. Further comparisons between L1 and L2 speakers are needed in order to assess whether this is true or not.

Conclusion

In this study, we have presented an approach to readability prediction that uses eye-tracking data collected from L1 and L2 speakers of English in order to induce text readability. Our multi-task learning models predict eye movement behavior as an auxiliary task in order to improve our main task.

Features	Precision	Recall	F1 score
	L1		
All Features	79.53	79.38	79.46
Ratio Verbs	68.50	98.13	80.00
Ratio Nouns	68.00	95.81	79.58
# NPs	67.88	95.27	79.28
Ave. word length	68.24	93.77	79.00
# AdvPs	67.39	97.00	79.53
# SBAR's	67.57	96.23	79.40
Ratio Adverbs	67.13	97.88	79.64
Parse tree height	67.13	97.88	79.64
# VPs	67.12	97.83	79.62
Ratio Pronouns	66.93	97.96	79.54
# PPs	67.56	96.23	79.39
Sentence length	68.23	94.00	78.99
Ratio Adjectives	67.52	96.11	79.32
Syn. surprisal Word1	69.70	88.28	77.90
Ambiguity word 1	69.24	86.11	76.76
Total surprisal	69.11	85.53	76.44
Lexical surprisal	69.18	85.03	76.29
	L2		
All Features	79.14	79.61	79.38
Ratio Nouns	67.36	95.73	79.07
Ratio Verbs	66.71	98.42	79.52
# NPs	67.39	95.33	78.96
Ave. word length	67.66	93.94	78.67
# SBAR's	66.64	96.87	78.66
# AdvPs	66.76	96.27	78.85
Ratio Adverbs	66.68	96.52	78.87
# VPs	66.62	96.24	78.74
Parse tree height	66.17	97.72	78.90
# PPs	65.82	98.60	78.94
Ratio Pronouns	65.83	98.03	78.77
Sentence length	67.26	91.72	77.61
Syntactic Surpisal	68.32	87.22	76.62

Table 5: Results from a single feature experiment using the Wikipedia dataset in which each feature was used to predict cases of improvement or loss between multi-task and single-task systems

First, we demonstrate how our method can be very beneficial for applications with small data samples as improvements of multi-task learning over single-task learning can be seen with as little as 100 training samples. In addition, we show that eye movement data from native speakers of English can be just as efficient as using data from L2 speakers. This is particularly beneficial for improving technologies aimed at language learners, since it may make data collection easier and less expensive.

Acknowledgement

Thanks to Maria Barrett, as well as our anonymous reviewers, for helpful advice. The second author was supported by ERC Starting Grant No. 313695.

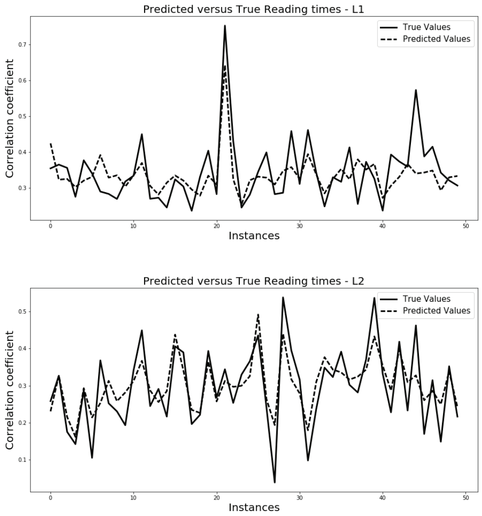


Figure 3: Plots showing the predictions of gaze made by our single-task MLP models versus the real values

References

Ambati, R. B.; Reddy, S.; and Steedman, M. 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*, 1051–1057.

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

Clifton, C.; Staub, A.; and Rayner, K. 2007. Eye movements in reading words and sentences. *Eye movements: A window on mind and brain* 27:341–372.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

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

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

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

Green, M. J. 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)*, 3846. Association for Computational Linguistics.

Jelinek, F., and Lafferty, J. D. 1991. Computation of the probability of initial substring generation by stochas-

<table>
<tr><th colspan="2">Improvements L1</th></tr>
<tr><td>Difficulty Level</td><td>Sentence</td></tr>
<tr><td>Normal</td><td>Mount Vesuvius is best known for its eruption in AD 79 that led to the destruction of the Roman cities of Pompeii and <b>Herculaneum and the death of 10,000 to 25,000 people.</b></td></tr>
<tr><td>Simplified</td><td>Pompeii was a Roman city; in 79 AD , a volcano called Mount Vesuvius erupted and destroyed the city and its people.</td></tr>
<tr><th colspan="2">Losses L1</th></tr>
<tr><td>Difficulty Level</td><td>Sentence</td></tr>
<tr><td>Normal</td><td>The Asociación Cultural de la Llingua Llionesa El Fueyu is a Leonese language association <b>whose main effort is promoting</b> the knowledge of Leonese language and the defense of the rights of Leonese language speakers .</td></tr>
<tr><td>Simplified</td><td>The Asociación Cultural de la Llingua Llionesa El Fueyu is a Leonese language <b>who wants to promote</b> the knowledge of Leonese language and the defense of the rights of Leonese language speakers.</td></tr>
<tr><th colspan="2">Improvements L2</th></tr>
<tr><td>Difficulty Level</td><td>Sentence</td></tr>
<tr><td>Normal</td><td><b>Consequently, the distance between the shock and the body generating it reduces</b> at high Mach numbers.</td></tr>
<tr><td>Simplified</td><td><b>Because of this, the shock layer is thin</b> at high mach numbers. .</td></tr>
<tr><th colspan="2">Losses L2</th></tr>
<tr><td>Difficulty Level</td><td>Sentence</td></tr>
<tr><td>Normal</td><td>In the far future it could grow into a long high mountain range if the continued northward movement of Africa <b>obliterates</b> the Mediterranean Sea .</td></tr>
<tr><td>Simplified</td><td>In the far future it could grow into a long and high mountain range if the continued northward movement of Africa <b>destroys</b> the Mediterranean Sea.</td></tr>
</table>

Table 6: This is an example of the sentences where we observed an improvement and a loss using multi-task over single-task for both L1 and L2 subsets. The bold text shows where the sentences differ. The first example however, relies more on summarization and compression.

tic context-free grammars. *Computational Linguistics* 17(3):315–323.

Kennedy, A., and Pynte, J. 2003. The dundee corpus. *Proceedings of the 12th European conference on eye movement*.

Klerke, S.; Castilho, S.; Barrett, M.; and Søgaard, A. 2015. Reading metrics for estimating task efficiency with mt output. In *Conference on Empirical Methods in Natural Language Processing*, 6.

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

Rayner, K.; Pollatsek, A.; Ashby, J.; and Clifton Jr, C. 2012. *Psychology of reading*. Psychology Press.

Roark, B.; Bachrach, A.; Cardenas, C.; and Pallier, C. 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*, 324333. Association for Computational Linguistics.

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

Singh, A. D.; Mehta, P.; Husain, S.; and Rajkumar, R. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, 202212.

Vajjala, S., and Meurers, D. 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*, 163173.

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

Xia, M.; Kochmar, E.; and Briscoe, E. 2016. Text readability assessment for second language learners.