Sequence classification with human attention

Maria Barrett[1] **Joachim Bingel**[2]
Nora Hollenstein[3] **Marek Rei**[4] **Anders Søgaard**[2]
[1]Department of Nordic Studies and Linguistics, University of Copenhagen, Denmark
[2]Department of Computer Science, University of Copenhagen, Denmark
[3]Department of Computer Science, ETH Zurich, Switzerland
[4]Department of Computer Science and Technology, University of Cambridge, United Kingdom
barrett@hum.ku.dk {bingel, soegaard}@di.ku.dk
noraho@ethz.ch marek.rei@cl.cam.ac.uk

Abstract

Learning attention functions requires large volumes of data, but many NLP tasks simulate human behavior, and in this paper, we show that human attention really does provide a good inductive bias on many attention functions in NLP. Specifically, we use estimated human attention derived from eye-tracking corpora to regularize attention functions in recurrent neural networks. We show substantial improvements across a range of tasks, including sentiment analysis, grammatical error detection, and detection of abusive language.

1 Introduction

When humans read a text, they do not attend to *all* its words (Carpenter and Just, 1983; Rayner and Duffy, 1988). For example, humans are likely to omit many function words and other words that are predictable in context and focus on less predictable content words. Moreover, when they fixate on a word, the duration of that fixation depends on a number of linguistic factors (Clifton et al., 2007; Demberg and Keller, 2008).

Since learning good attention functions for recurrent neural networks requires large volumes of data (Zoph et al., 2016; Britz et al., 2017), and errors in attention are known to propagate to classification decisions (Alkhouli et al., 2016), we explore the idea of using human attention, as estimated from eye-tracking corpora, as an inductive bias on such attention functions. Penalizing attention functions for departing from human attention may enable us to learn better attention functions when data is limited.

Eye-trackers provide millisecond-accurate records on where humans look when they are reading, and they are becoming cheaper and more easily available by the day (San Agustin et al., 2009). In this paper, we use publicly available eye-tracking corpora, i.e., texts augmented with eye-tracking measures such as fixation duration times, and large eye-tracking corpora have appeared increasingly over the past years. Some studies suggest that the relevance of text can be inferred from the gaze pattern of the reader (Salojärvi et al., 2003) – even on word-level (Loboda et al., 2011).

Contributions We present a recurrent neural architecture with attention for sequence classification tasks. The architecture jointly learns its parameters and an attention function, but can alternate between supervision signals from labeled sequences (with no explicit supervision of the attention function) and from attention trajectories. This enables us to use per-word fixation durations from eye-tracking corpora to regularize attention functions for sequence classification tasks. We show such regularization leads to significant improvements across a range of tasks, including sentiment analysis, detection of abusive language, and grammatical error detection. Our implementation is made available at https://github.com/coastalcph/Sequence_classification_with_human_attention.

2 Method

We present a recurrent neural architecture that jointly learns the recurrent parameters and the attention function, but can alternate between supervision signals from labeled sequences and from attention trajectories in eye-tracking corpora. The input will be a set of labeled sequences (sentences paired with discrete category labels) and a set of sequences, in which each token is associated with a scalar value representing the attention human readers devoted to this token on average.

The two input datasets, i.e., the target task train-

Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018), pages 302–312
Brussels, Belgium, October 31 - November 1, 2018. ©2018 Association for Computational Linguistics

ing data of sentences paired with discrete categories, and the eye-tracking corpus, need not (and will not in our experiments) overlap in any way. Our experimental protocol, in other words, does not require in-task eye-tracking recordings, but simply leverages information from existing, available corpora.

Behind our approach lies the simple observation that we can correlate the token-level attention devoted by a recurrent neural network, even if trained on sentence-level signals, with any measure defined at the token level. In other words, we can compare the attention devoted by a recurrent neural network to various measures, including token-level annotation (Rei and Søgaard, 2018) and eye-tracking measures. The latter is particularly interesting as it is typically considered a measurement of *human* attention.

We go beyond this: Not only can we compare machine attention with human attention, we can also constrain or inform machine attention by human attention in various ways. In this paper, we explore this idea, proposing a particular architecture and training method that, in effect, uses human attention to *regularize* machine attention.

Our training method is similar to a standard approach to training multi-task architectures (Dong et al., 2015; Søgaard and Goldberg, 2016; Bingel and Søgaard, 2017), sometimes referred to as the *alternating* training approach (Luong et al., 2016): We randomly select a data point from our training data or the eye-tracking corpus with some (potentially equal) probability. If the data point is sampled from our training data, we predict a discrete category and use the computed loss to update our parameters. If the data point is sampled from the eye-tracking corpus, we still run the recurrent network to produce a category, but this time we only monitor the attention weights assigned to the input tokens. We then compute the minimum squared error between the normalized eye-tracking measure and the normalized attention score. In other words, in multi-task learning, we optimize each task for a fixed number of parameter updates (or mini-batches) before switching to the next task (Dong et al., 2015); in our case, we optimize for a target task (for a fixed number of updates), then improve our attention function based on human attention (for a fixed number of updates), then return to optimizing for the target task and continue iterating.

2.1 Model

Our architecture is a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) that encodes word representations x_i into forward and backward representations, and into combined hidden states h_i (of slightly lower dimensionality) at every timestep. In fact, our model is a hierarchical model whose word representations are concatenations of the output of character-level LSTMs and word embeddings, following Plank et al. (2016), but we ignore the character-level part of our architecture in the equations below:

$$\overrightarrow{h_i} = LSTM(x_i, \overrightarrow{h_{i-1}}) \tag{1}$$

$$\overleftarrow{h_i} = LSTM(x_i, \overleftarrow{h_{i+1}}) \tag{2}$$

$$\widetilde{h_i} = [\overrightarrow{h_i}; \overleftarrow{h_i}] \tag{3}$$

$$h_i = \tanh(W_h \widetilde{h_i} + b_h) \tag{4}$$

The final (reduced) hidden state is sometimes used as a sentence representation s, but we instead use attention to compute s by multiplying dynamically predicted attention weights with the hidden states for each time step. The final sentence predictions y are then computed by passing s through two more hidden layers:

$$s = \sum_i \widetilde{a_i} h_i \tag{5}$$

$$y = \sigma(W_y \tanh(W_{\tilde{y}} s + b_{\tilde{y}}) + b_y) \tag{6}$$

From the hidden states, we directly predict token-level raw attention scores a_i:

$$e_i = \tanh(W_e h_i + b_e) \tag{7}$$

$$a_i = W_a e_i + b_a \tag{8}$$

We normalize these predictions to attention weights $\widetilde{a_i}$:

$$\widetilde{a_i} = \frac{a_i}{\sum_k a_k} \tag{9}$$

Our model thus combines two distinct objectives: one at the sentence level and one at the token level. The sentence-level objective is to minimize the squared error between output activations and true sentence labels $\widehat{y}$.

$$L_{sent} = \sum_j (y^{(j)} - \widehat{y}^{(j)})^2 \tag{10}$$

The token-level objective, similarly, is to minimize the squared error for the attention not aligning with our human attention metric.

$$L_{tok} = \sum_j \sum_t (a^{(j)(t)} - \widehat{a}^{(j)(t)})^2 \qquad (11)$$

These are finally combined to a weighted sum, using λ (between 0 and 1) to trade off loss functions at the sentence and token levels.

$$L = L_{sent} + \lambda L_{tok} \qquad (12)$$

Note again that our architecture does not require the target task data to come with eye-tracking information. We instead learn jointly to predict sentence categories and to attend to the tokens humans tend to focus on for longer. This requires a training schedule that determines when to optimize for the sentence-level classification objective, and when to optimize the machine attention at the token level. We therefore define an epoch to comprise a fixed number of batches, and sample every batch of training examples either from the target task data or from the eye-tracking corpus, as determined by a coin flip, the bias of which is tuned as a hyperparameter. Specifically, we define an epoch to consist of n batches, where n is the number of training sentences in the target task data divided by the batch size. This coin is potentially weighted with data being drawn from the auxiliary task with some probability or a decreasing probability of $\frac{1}{E+1}$, where E is the current epoch; see Section 4 for hyper-parameters.

3 Data

As mentioned in the above, our architecture requires no overlap between the eye-tracking corpus and the training data for the target task. We therefore rely on publicly available eye-tracking corpora. For sentiment analysis, grammatical error detection, and hate speech detection, we use publicly available research datasets that have been used previously in the literature. All datasets were lower-cased.

3.1 Eye-tracking corpora

For our experiments, we concatenate two publicly available eye-tracking corpora, the Dundee Corpus (Kennedy et al., 2003) and the reading parts of the ZuCo Corpus (Hollenstein et al., 2018), described below. Both corpora contain eye-tracking measurements from several subjects reading the same text. For every token, we compute the mean duration of all fixations to this token as our measure of human attention, following previous work (Barrett et al., 2016a; Gonzalez-Garduno and Søgaard, 2018).

Dundee The English part of the Dundee corpus (Kennedy et al., 2003) comprises 2,368 sentences and more than 50,000 tokens. The texts were read by ten skilled, adult, native speakers. The texts are 20 newspaper articles from *The Independent*. The reading was self-paced and as close to natural, contextualized reading as possible for a laboratory data collection. The apparatus was a Dr Bouis Oculometer Eyetracker with a 1000 Hz monocular (right) sampling. At most five lines were shown per screen while subjects were reading.

ZuCo The ZuCo corpus (Hollenstein et al., 2018) is a combined eye-tracking and EEG dataset. It contains approximately 1,000 individual English sentences read by 12 adult, native speakers. Eye movements were recorded with the infrared video-based eye tracker *EyeLink 1000 Plus* at a sampling rate of 500 Hz. The sentences were presented at the same position on the screen, one at a time. Longer sentences spanned multiple lines. The subjects used a control pad to switch to the next sentence and to answer the control questions, which allowed for natural reading speed. The corpus contains both natural reading and reading in a task-solving context. For compatibility with the Dundee corpus, we only use the subset of the data, where humans were encouraged to read more naturally. This subset contains 700 sentences. This part of the Zuco corpus contains positive, negative or neutral sentences from the Stanford Sentiment Treebank (Socher et al., 2013) for passive reading, to analyze the elicitation of emotions and opinions during reading. As a control condition, the subjects sometimes had to rate the quality of the described movies; in approximately 10% of the cases. The Zuco corpus also contains instances where subjects were presented with Wikipedia sentences that contained semantic relations such as *employer*, *award* and *job_title* (Culotta et al., 2006). The control condition for this tasks consisted of multiple-choice questions about the content of the previous sentence; again, approximately 10% of all sentences were followed by a question.

Task	Trains set		Dev. set		Test set	
	domain	n sent	domain	n sent	domain	n sent
Sentiment	SemEval Twitter	7,177	SemEval Twitter	1,205	SemEval Twitter	2,870
Sentiment					SemEval SMS	2,094
Grammatical error	FCE	28,731	FCE	2,222	FCE	2,720
Abusive language	Waseem (2016)	5,529	Waseem (2016)	690	Waseem (2016)	690
Abusive language	Waseem and Hovy (2016)	11,225	Waseem and Hovy (2016)	1,403	Waseem and Hovy (2016)	1,403

Table 1: Overview of the tasks and datasets used.

Preprocessing of eye-tracking data Mean fixation duration (Mean fix dur) is extracted from the Dundee Corpus. For Zuco, we divide total reading time per word token with the number of fixations to obtain mean fixation duration. The mean fixation duration is selected empirically among gaze duration (sum of all fixations in the first pass reading of the a word) and total fixation duration, and n fixations. Then we average these numbers for all readers of the corpus to get a more robust average processing time. Eye-tracking is known to correlate with word frequency (Rayner and Duffy, 1988). We include a frequency baseline on the eye tracking text, BNC inv freq. The word frequencies comes from the British National Corpus (BNC) frequency lists (Kilgarriff, 1995). We use log-transformed frequency per million. Before normalizing, we take the additive inverse of the frequency, such that rare words get a high value, making it comparable to gaze.

Mean fix dur and BNC inv freq are min-max-normalized to a value in the range 0-1. Mean fix dur is normalized separately for the two eye tracking corpora. We expect the experimental bias – especially the fact that ZuCo contains reading of isolated sentences and Dundee contains longer texts – to influence the reading and therefore separate normalization should preserve the signal within each corpus better.

3.2 Sentiment classification

Table 1 presents an overview of all train, development and test sets used in this paper.

Our first task is sentence-level sentiment classification. We note that many sentiment analysis datasets contain document-level labels or include more fine-grained annotation of text spans, say phrases or words. For compatibility with our other tasks, we focus on sentence-level sentiment analysis. We use the SemEval-2013 Twitter dataset (Wilson et al., 2013; Rosenthal et al., 2015) for training and development. For test, we use a same-domain test set, the SemEval-2013 Twitter test set (Semeval Twitter POS | NEG), and an out-of-domain test set, SemEval-2013 SMS test set (Semeval SMS POS | NEG). The SemEval-2013 sentiment classification task was a three-way classification task with positive, negative and neutral classes. We reduce the task to binary tasks detecting negative sentences vs. non-negative and vice versa for the positive class. Therefore the dataset size is the same for POS and NEG experiments.

3.3 Grammatical error detection

Our second task is grammatical error detection. We use the First Certificate in English error detection dataset (FCE) (Yannakoudakis et al., 2011). This dataset contains essays written by English learners during language examinations, where any grammatical errors have been manually annotated by experts. Rei and Yannakoudakis (2016) converted the dataset for a sequence labeling task and we use their splits for training, development and testing. Similarly to Rei and Søgaard (2018), we perform sentence-level binary classification of sentences that need some editing vs. grammatically correct sentences. We do not use the token-level labels for training our model.

3.4 Hate speech detection

Our third and final task is detection of abusive language; or more specifically, hate speech detection. We use the datasets of Waseem (2016) and Waseem and Hovy (2016). The former contains 6,909 tweets; the latter 14,031 tweets. They are manually annotated for sexism and racism. In this study, sexism and racism are conflated into one category in both datasets. Both datasets are split in train, development and test splits consisting of 80%, 10% and 10% of the tweets respectively.

4 Experiments

Models In our experiments, we compare three models: (a) a baseline model with automatically learned attention, (b) our model with an attention

	BL			BNC inv freq			Mean fix dur		
task	P	R	F_1	P	R	F_1	P	R	F_1
SemEval SMS neg	43.55	45.41	43.77	45.82	48.65	45.24	47.15	46.98	**45.77**
SemEval SMS pos	65.79	50.81	57.08	65.92	51.04	57.45	65.46	52.95	**58.50**
SemEval Twitter neg	57.39	26.87	35.70	62.50	28.66	37.78	60.52	30.67	**40.23**
SemEval Twitter pos	77.96	53.88	63.63	79.66	54.66	64.78	78.77	55.35	**64.96**
FCE	79.01	89.33	83.84	79.18	89.26	83.89	79.03	90.28	**84.28**
Waseem (2016)	76.42	62.07	68.29	77.20	61.71	68.54	77.20	63.06	**69.30**
Waseem and Hovy (2016)	76.23	72.23	74.16	76.33	74.70	75.48	76.95	74.43	**75.61**
Mean	68.05	57.23	60.92	69.52	58.38	61.88	69.30	59.10	**62.67**

Table 2: Sentence classification results. P(recision), R(ecall) and F_1. Averages over 10 random seeds. The best average F_1 score per task is shown in bold.

function regularized by information about human attention, and finally, (c) a second baseline using frequency information as a proxy for human attention and using the same regularization scheme as in our human attention model.

Hyperparameters Basic hyper-parameters such as number of hidden layers, layer size, and activation functions were following the settings of Rei and Søgaard (2018). The dimensionality of our word embedding layer was set to size 300, and we use publicly available pre-trained Glove word embeddings (Pennington et al., 2014) that we finetune during training. The dimensionality of the character embedding layer was set to 100. The recurrent layers in the character-level component have dimensionality 100; the word-level recurrent layers dimensionality 300. The dimensionality of our feed-forward layer, leading to reduced combined representations h_i, is 200, and the attention layer has dimensionality 100.

Three hyper-parameters, however, we tune for each architecture and for each task, by measuring sentence-level F_1-scores on the development sets. These are: (a) learning rate, (b) λ in Equation (12), i.e., controlling the relative importance of the attention regularization, and (c) the probability of sampling data from the eye-tracking corpus during training.

For all tasks and all conditions (baseline, frequency-informed baseline, and our human attention model), we perform a grid search over learning rates [.01 .1 1.], L_{att} weight λ values [.2 .4 .6 .8 1.], and probability of sampling from the eye-tracking corpus [.125 .25 .5 1., decreasing] – where *decreasing* means that the probability of sampling from the eye-tracking corpus initially is 0.5, but drops linearly for each epoch ($\frac{1}{E+1}$; see 2.1. We apply the models with the best average F_1 scores over three random seeds on the validation data, to our test sets.

Initialization Our models are randomly initialized. This leads to some variance in performance across different runs. We therefore report averages over 10 runs in our experiments below.

5 Results

Our performance metric across all our experiments is the sentence-level F_1 score. We report precision, recall and F_1 scores for all tasks in Table 2.

Our main finding is that our human attention model, based on regularization from mean fixation durations in publicly available eye-tracking corpora, consistently outperforms the recurrent architecture with learned attention functions. The improvements over both baseline and BNC frequency are significant ($p < 0.01$) using bootstrapping (Calmettes et al., 2012) over all tasks, with one seed. The mean error reduction over the baseline is 4.5%.

Unsurprisingly, knowing that human attention helps guide our recurrent architecture, the frequency-informed baseline is also better than the non-informed baseline across the board, but the human attention model is still significantly better across all tasks ($p < 0.01$). For all tasks except negative sentiment, we note that generally, most of the improvements over the learned attention baseline for the gaze-informed models, are due to improvements in recall. Precision is not worse, but we do not see any larger improvements on precision either. For the negative SEMEVAL tasks, we also see larger improvements for precision.

The observation that improvements are primarily due to increased recall, aligns well with the hypothesis that human attention serves as an efficient regularization, preventing overfitting to surface statistical regularities that can lead the network to rely on features that are not there at test time (Globerson and Roweis, 2006), at the expense of target class precision.

6 Analysis

We illustrate the differences between our baseline models and the model with gaze-informed attention by the attention weights of an example sentence. Though it is a single, cherry-picked example, it is representative of the general trends we observe in the data, when manually inspecting attention patterns. Table 3 presents a coarse visualization of the attention weights of six different models, namely our baseline architecture and the architecture with gaze-informed attention, trained on three different tasks: hate speech detection, negative sentiment classification, and error detection. The sentence is a positive hate speech example from the Waseem and Hovy (2016) development set. The words with more attention than the sentence average are bold-faced.

First note that the baseline models only attend to one or two coherent text parts. This pattern was very consistent across all the sentences we examined. This pattern was not observed with gaze-informed attention.

Our second observation is that the baseline models are more likely to attend to stop words than gaze-informed attention. This suggests that gaze-informed attention has learned to simulate human attention to some degree. We also see many differences between the jointly learned task-specific, gaze-informed attention functions.

The gaze-informed hate speech classifier, for example, places considerable attention on *BUT*, which in this case is a passive-aggressive hate speech indicator. It also gives weight to *double standards* and *certain rules*.

The gaze-informed sentiment classifier, on the other hand, focuses more on *sorry I am not sexist* which, in isolation, reads like an apologetic disclaimer. This model also gives weight to *double standards* and *certain rules*

The gaze-informed grammatical error detection model gives attention to *standards*, which is ungrammatical, because of the morphological number disagreement with its determiner *a*; it also gives attention to *certain rules*, which is disagreeing, again in number, with *there's*. It also gives attention to the non-word *fem*.

Overall, this, in combination with our results in Table 3, suggests that the regularization effect from human attention enables our architecture to learn to better attend to the most relevant aspects of sentences for the target tasks. In other words, human attention provides the inductive bias that makes learning possible.

7 Discussion and related work

Gaze in NLP It has previously been shown that several NLP tasks benefit from gaze information, including part-of-speech tagging (Barrett and Søgaard, 2015b; Barrett et al., 2016a), prediction of multiword expressions (Rohanian et al., 2017) and sentiment analysis (Mishra et al., 2017b).

Gaze information and other measures from psycholinguistics have been used in different ways in NLP. Some authors have used discretized, single features (Pate and Goldwater, 2011, 2013; Plank, 2016; Klerke et al., 2016), whereas others have used multidimensional, continuous values (Barrett et al., 2016a; Bingel et al., 2016). We follow Gonzalez-Garduno and Søgaard (2018) in using a single, continuous feature. We did not experiment with other representations, however. Specifically, we only considered the signal from token-level, normalized mean fixation durations.

Fixation duration is a feature that carries an enormous amount of information about the text and the language understanding process. Carpenter and Just (1983) show that readers are more likely to fixate on open-class words that are not predictable from context, and Kliegl et al. (2004) show that a higher cognitive load results in longer fixation durations. Fixations before skipped words are shorter before short or high-frequency words and longer before long or low-frequency words in comparison with control fixations (Kliegl and Engbert, 2005). Many of these findings suggest correlations with syntactic information, and many authors have confirmed that gaze information is useful to discriminate between syntactic phenomena (Demberg and Keller, 2008; Barrett and Søgaard, 2015a,b).

Gaze data has also been used in the context of

FCE		SemEval Twitter NEG		Waseem and Hovy (2016)	
Bl	MFD	Bl	MFD	Bl	MFD
@CharlesClassiqk:	**@CharlesClassiqk:**	**@CharlesClassiqk:**	**@CharlesClaqqqqqqqssiqk:**	@CharlesClassiqk:	@CharlesClassiqk:
sorry	**sorry**	**sorry**	**sorry**	sorry	sorry
I'm	I'm	**I'm**	**I'm**	I'm	I'm
not	not	**not**	**not**	not	not
sexist	**sexist**	**sexist**	**sexist**	sexist	sexist
BUT	BUT	BUT	BUT	BUT	**BUT**
there	**there**	there	there	there	**there**
is	is	is	is	is	is
a	a	a	a	a	a
double	**double**	double	**double**	double	**double**
standards	**standards**	standards	**standards**	**standards**	**standards**
there's	**there's**	there's	there's	**there's**	**there's**
certain	**certain**	certain	**certain**	**certain**	**certain**
rules	rules	rules	**rules**	**rules**	**rules**
for	for	for	for	**for**	**for**
dudes	dudes	dudes	dudes	**dudes**	**dudes**
and	and	and	and	**and**	and
there's	**there's**	there's	there's	**there's**	**there's**
certain	**certain**	certain	**certain**	**certain**	**certain**
rules	**rules**	rules	**rules**	**rules**	**rules**
for	for	for	for	**for**	for
femâĂę	**femâĂę**	femâĂę	femâĂę	**femâĂę**	femâĂę

Table 3: One sentence marked as containing sexism from Waseem and Hovy (2016) development set. Using trained baseline (Bl) and gaze model (MFD) for three tasks: error detection, sentiment classification, and hate speech detection. Words with more attention than sentence average are boldfaced.

sentiment analysis before (Mishra et al., 2017b,a). Mishra et al. (2017b) augmented a sentiment analysis system with eye-tracking features, including first fixation durations and fixation counts. They show that fixations not only have an impact in detecting sentiment, but also improve sarcasm detection. They train a convolutional neural network that learns features from both gaze and text and uses them to classify the input text (Mishra et al., 2017a). On a related note, Raudonis et al. (2013) developed a emotion recognition system from visual stimulus (not text) and showed that features such as pupil size and motion speed are relevant to accurately detect emotions from eye-tracking data. Wang et al. (2017) use variables shown to correlate with human attention, e.g. surprisal, to guide the attention for sentence representations.

Gaze has also been used in the context of grammaticality (Klerke et al., 2015a,b), as well as in readability assessment (Gonzalez-Garduno and Søgaard, 2018).

Gaze has either been used as features (Barrett and Søgaard, 2015a; Barrett et al., 2016b) or as a direct supervision signal in multi-task learning scenarios (Klerke et al., 2016; Gonzalez-Garduno and Søgaard, 2018). We are, to the best of our knowledge, the first to use gaze to inform attention functions in recurrent neural networks.

Human-inspired attention functions Ibraheem et al. (2017), however, uses optimal attention to simulate human attention in an interactive machine translation scenario, and Britz et al. (2017) limit attention to a local context, inspired by findings in studies of human reading. Rei and Søgaard (2018) use auxiliary data to regularize attention functions in recurrent neural networks; not from psycholinguistics data, but using small amounts of task-specific, token-level annotations. While their motivation is very different from ours, technically our models are very related. In a different context, Das et al. (2017) investigated whether humans attend to the same regions as neural networks solving visual question answering problems. Lindsey (2017) also used human-inspired, unsupervised attention in a computer vision context.

Other work on multi-purpose attention functions While our work is the first to use gaze data to guide attention in a recurrent architectures, there has recently been some work on sharing attention functions across tasks. Firat et al. (2016), for example, share attention functions between languages in the context of multi-way neural machine translation.

Sentiment analysis While sentiment analysis is most often considered a supervised learning problem, several authors have leveraged other signals than annotated data to learn sentiment analysis models that generalize better. Felbo et al. (2017), for example, use emoji prediction to pretrain their sentiment analysis models. Mishra et al. (2018) use several auxiliary tasks, including gaze prediction, for document-level sentiment analysis. There is a lot of previous work, also, leveraging information across different sentiment analysis datasets, e.g., Liu et al. (2016).

Error detection In grammatical error detection, Rei (2017) used an unsupervised auxiliary language modeling task, which is similar in spirit to our second baseline, using frequency information as auxiliary data. Rei and Yannakoudakis (2017) go beyond this and evaluate the usefulness of many auxiliary tasks, primarily syntactic ones. They also use frequency information as an auxiliary task.

Hate speech detection In hate speech detection, many signals beyond the text are often leveraged (see Schmidt and Wiegand (2017) for an overview of the literature). Interestingly, many authors have used signals from sentiment analysis, e.g., Gitari et al. (2015), motivated by the correlation between hate speech and negative sentiment. This correlation may also explain why we see the biggest improvements with gaze-informed attention on those two tasks.

Human inductive bias Finally, our work relates to other work on providing better inductive biases for learning human-related tasks by observing humans (Tamuz et al., 2011; Wilson et al., 2015). We believe this is a truly exciting line of research that can help us push research horizons in many ways.

8 Conclusion

We have shown that human attention provides a useful inductive bias on machine attention in recurrent neural networks for sequence classification problems. We present an architecture that enables us to leverage human attention signals from general, publicly available eye-tracking corpora, to induce better, more robust task-specific NLP models. We evaluate our architecture and show improvements across three NLP tasks, namely sentiment analysis, grammatical error detection, and detection of abusive language. We observe that not only does human attention help models distribute their attention in a generally useful way; human attention also seems to act like a regularizer providing more robust performance across domains, and it enables better learning of task-specific attention functions through joint learning.

References

Tamer Alkhouli, Gabriel Bretschner, Jan-Thorsten Peter, Mohammed Hethnawi, Andreas Guta, and Hermann Ney. 2016. Alignment-based neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 54–65.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016a. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pages 579–584.

Maria Barrett, Frank Keller, and Anders Søgaard. 2016b. Cross-lingual transfer of correlations between parts of speech and gaze features. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 1330–1339.

Maria Barrett and Anders Søgaard. 2015a. Reading behavior predicts syntactic categories. In *Proceedings of the nineteenth conference on computational natural language learning (CoNLL)*, pages 345–249.

Maria Barrett and Anders Søgaard. 2015b. Using reading behavior to predict grammatical functions. In *Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*, pages 1–5.

Joachim Bingel, Maria Barrett, and Anders Søgaard. 2016. Extracting token-level signals of syntactic processing from fMRI-with an application to POS induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 747–755.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, volume 2, pages 164–169.

Denny Britz, Melody Y. Guan, and Minh-Thang Luong. 2017. Efficient attention using a fixed-size memory representation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 392–400.

Guillaume Calmettes, Gordon B Drummond, and Sarah L Vowler. 2012. Making do with what we have: use your bootstraps. *The Journal of physiology*, 590(15):3403–3406.

Patricia A Carpenter and Marcel Adam Just. 1983. What your eyes do while your mind is reading. *Eye movements in reading: Perceptual and language processes*, pages 275–307.

Charles Clifton, Adrian Staub, and Keith Rayner. 2007. Eye movements in reading words and sentences. In *Eye Movements: A Window on Mind and Brain*, pages 341–371. Elsevier, Amsterdam, The Netherlands.

Aron Culotta, Andrew McCallum, and Jonathan Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 296–303. Association for Computational Linguistics.

Abhishek Das, Harsh Agrawal, Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1723–1732.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyan Rahwan, and Sune. Lehmann. 2017. Using millions of emoji occurrences to pretrain any-domain models for detecting emotion, sentiment, and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 866–875.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Amir Globerson and Sam Roweis. 2006. Nightmare at test time: robust learning by feature deletion. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 353–360.

Ana Gonzalez-Garduno and Anders Søgaard. 2018. Learning to predict readability using eye-movement data from natives and learners. In *Proceedings of the Thirty-Second Association for the Advancement of Artificial Intelligence Conference (AAAI)*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo: A simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific data, Under Review*.

Samee Ibraheem, Nicholas Altieri, and John DeNero. 2017. Learning an interactive attention policy for neural machine translation. In *MTSummit*.

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

Adam Kilgarriff. 1995. BNC database and word frequency lists. *Retrieved Dec. 2017*.

Sigrid Klerke, Héctor Martínez Alonso, and Anders Søgaard. 2015a. Looking hard: Eye tracking for detecting grammaticality of automatically compressed sentences. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 97–105.

Sigrid Klerke, Sheila Castilho, Maria Barrett, and Anders Søgaard. 2015b. Reading metrics for estimating task efficiency with MT output. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*, pages 6–13.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1528–1533.

Reinhold Kliegl and Ralf Engbert. 2005. Fixation durations before word skipping in reading. *Psychonomic Bulletin & Review*, 12(1):132–138.

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(1-2):262–284.

Jack Lindsey. 2017. Pre-training attention mechanisms. In *NIPS Workshop on Cognitive Informed Artificial Intelligence*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Deep multi-task learning with shared memory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 118–127.

Tomasz D Loboda, Peter Brusilovsky, and Jöerg Brunstein. 2011. Inferring word relevance from eye-movements of readers. In *Proceedings of the 16th international conference on intelligent user interfaces*, pages 175–184. ACM.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence-to-sequence learning. In *International Conference on Learning Representations (ICLR)*.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017a. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 377–387.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2017b. Leveraging cognitive features for sentiment analysis. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 156–166.

Abhijit Mishra, Srikanth Tamilselvam, Riddhiman Dasgupta, Seema Nagar, and Kuntal Dey. 2018. Cognition-cognizant sentiment analysis with multitask subjectivity summarization based on annotators' gaze behavior. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*.

John K Pate and Sharon Goldwater. 2011. Unsupervised syntactic chunking with acoustic cues: computational models for prosodic bootstrapping. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 20–29.

John K Pate and Sharon Goldwater. 2013. Unsupervised dependency parsing with acoustic cues. *Transactions of the Association for Computational Linguistics (TACL)*, 1:63–74.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

Barbara Plank, Yoav Goldberg, and Anders Søgaard. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 412–418.

Vidas Raudonis, Gintaras Dervinis, Andrius Vilkauskas, Agne Paulauskaite-Taraseviciene, and Gintare Kersulyte-Raudone. 2013. Evaluation of human emotion from eye motions. *Evaluation*, 4(8).

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, NY, USA. Academic Press.

Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 2121–2130.

Marek Rei and Anders Søgaard. 2018. Zero-shot sequence labeling: Transferring knowledge from sentences to tokens. *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*, pages 293–302.

Marek Rei and Helen Yannakoudakis. 2016. Compositional sequence labeling models for error detection in learner writing. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1181–1191.

Marek Rei and Helen Yannakoudakis. 2017. Auxiliary objectives for neural error detection models. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 33–43.

Omid Rohanian, Shiva Taslimipoor, Victoria Yaneva, and Le An Ha. 2017. Using gaze data to predict multiword expressions. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 601–609.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 Task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 451–463.

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, John Paulin Hansen, and Dan Witzner Hansen. 2009. Low-cost gaze interaction: ready to deliver the promises. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 4453–4458. ACM.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 231–235.

Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. 2011. Adaptively learning the crowd kernel. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 673–680.

Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2017. Learning sentence representation with guidance of human attention. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 4137–4143.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Andrew Wilson, Christoph Dann, Chris Lucas, and Eric Xing. 2015. The human kernel. In *Advances in neural information processing systems (NIPS)*, pages 2854–2862.

Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter. 2013. Sentiment analysis in Twitter. In *Proceedings of the Seventh International Workshop on Semantic*, pages 312–320.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 180–189. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1568–1575.