Abstractive Automatic Text Summarization with Eye-tracking Data

Xuemin Duan
University of Copenhagen
npm838@alumni.ku.dk

Abstract

Eye-tracking data show great potential in improving Natural Language Processing tasks. This paper explores whether gaze features can improve the text generation performance on Automatic Text Summarization. Specifically, I implemented a bidirectional LSTM model with different types of word embeddings, and the experimental results show that the word embeddings combining ZuCo and GECO outperform the baseline model by 1% of the F_1 score of ROUGE-1. I also apply an ANOVA test to find out whether there is a statistically significant difference between the results of the models and which groups of results are different from each other. It turns out that there are significant differences between the results of the baseline model and that of the models with eye-tracking features, but no significant difference between the results of the eye-tracking feature models. The codes have been released in https://github.com/xuemduan/CogSci3.

1 Introduction

Eye-tracking data can quantify humans' visual and cognitive attention since it measures the point of gaze and eye movement relative to the head at a given time, objectively monitoring where, when and what people are looking at. Natural language processing (NLP) focuses on tasks related to reading, writing, and speaking which humans cognitively perform every day. The gaze features show great potential in improving NLP tasks.

(Hollenstein et al., 2019a) employed gaze features on Named Entity Recognition, Relation Classification, and Sentiment Analysis tasks and received the results which exceeded the baselines. (Barrett et al., 2016) used eye-tracking data to improve the Part-of-speech Tagging task and also get a result that significantly outperforms the baseline without gaze features. These tasks can be regarded as the classification tasks of NLP.

With the explosive growth of text information, NLP researchers show an increased interest in ATS which can be divided into extractive and abstractive. The extractive ATS has relatively stable performance but its upper bound is very limited since it can only select sentences from the original article to combine the candidate summary. The abstractive ATS is more difficult but has greater potential to approach the level of the manual summary because it can generate text flexibly which is consistent with what humans do when summarizing the article. However, most studies in the field of using eye-tracking data to improve NLP tasks have only focused on classification task but are limited to text generation, especially on Automatic Text Summarization (ATS).

This paper aims to contribute to this research gap by exploring the performance of eye-tracking data on improving abstractive ATS. With my work, the contributions are as follows:

- I implemented a bidirectional LSTM model with different word embeddings, and the results show that the eye-tracking data can significantly improve the text generation performance on abstractive ATS
- I analyzed whether there is a statistically significant difference between the results of the models, and find out exactly which groups of results are different from each other.
- I discussed the limitations of applying eye-tracking data on abstractive Automatic Text Summarization.

2 Related Work

Eye-tracking data can be effectively used to evaluate the model performance on NLP tasks. (Green, 2014) and (Hahn and Keller, 2016) argued that eye-tracking data recorded from reading can well explain human reading comprehension and can

be used in Natural Language Generation (NLG). (Klerke et al., 2015) showed that the gaze features reflect aspects of reading through sentence comprehension task, which is helpful for the evaluation of automatic text summarization.

Eye-tracking data can also effectively improve some NLP tasks. (Hollenstein et al., 2019a) employed the gaze and EEG features on Named Entity Recognition, Relation Classification, and Sentiment Analysis tasks and significantly outperform the baselines. (Barrett et al., 2016) successfully improved the Part-of-speech Tagging task with gaze features.

(Hollenstein et al., 2018) present a dataset combining eye-tracking and electroencephalography (EEG) for English. Ghent EyeTracking Corpus (GECO) (Cop et al., 2017) is a freely available English monolingual and Dutch–English bilingual eye-tracking dataset. (Yi et al., 2020) present a dataset for exploring gaze behaviors in Chinese text summarization and provides the connection between humans' reading behaviors and the summaries written by them.

3 Dataset

This paper implemented ATS on the Amazon Fine Food Reviews dataset, and used gaze features from ZuCo and GECO to improve the performance.

3.1 Amazon Fine Food Reviews

Amazon Fine Food Reviews (McAuley and Leskovec, 2013) is a freely available dataset consisting of reviews and corresponding summaries of fine foods from amazon. The data spans more than 10 years and includes about 500,000 reviews up to October 2012. After removing some exception examples, the dataset is left with 393,222 texts and summaries. The description of the dataset is shown in Figure 1. The x-axis defines the number of words in text and summary and the y-axis defines the number of texts and summaries. The average number of words contained in the summary is 4.0 and that in the text is 37.7.

3.2 Zurich Cognitive Language Processing Corpus (ZuCo)

Zurich Cognitive Language Processing Corpus (ZuCo) (Hollenstein et al., 2018) is a dataset combining eye-tracking and electroencephalography (EEG) which is recorded from participants reading sentences based on different tasks. There are

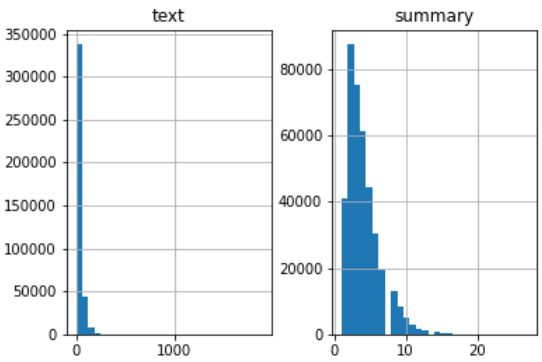

Figure 1: Amazon Fine Food Reviews dataset.

two versions, ZuCo 1.0 and ZuCo 2.0 (Hollenstein et al., 2019b).

ZuCo 1.0 It is recorded from 12 healthy adults. All the participants are native English speakers originating from Canada, USA, UK, and Australia, of which 5 females and 7 males, and are all right-handed and between 22 and 54 years old. The eye-tracking data were recorded using the EyeLink 1000 system while participants read English sentences which from the Stanford Sentiment Treebank (Socher et al., 2013) and the Wikipedia dataset (Culotta et al., 2006).

ZuCo 2.0 It is recorded from 19 healthy adults and finally discarded one. All the participants are native English speakers originating from Canada, USA, UK, Australia, and South Africa, of which 10 females and 8 males, two of them are left-handed, and three of them wear glasses. The eye-tracking data were also recorded using the EyeLink 1000 system while participants reading English sentences which from the Wikipedia corpus.

The studies for ZuCo 1.0 and ZuCo 2.0 were all approved by the Ethics Commission of the University of Zurich (Hollenstein et al., 2018). And all participants have written consent to participate.

The eye-tacking data from ZuCo used in this paper is the version released in Absalon of Cognitive Science 3 which was processed for the CMCL 2021 Shared Task[1]. This version extracted eye-tracking data from Task 1 and Task 2 from ZuCo 1.0 and all tasks from ZuCo 2.0. This paper combines its training set and test set into one set and plot its feature value distribution in Figure 2, finally coming up with 5 eye-tracking features of 6,259 unique words: nFix (number of fixations), FFD (first fixation duration), GPT (go-past time), TRT (total reading time),

[1] https://competitions.codalab.org/competitions/28176

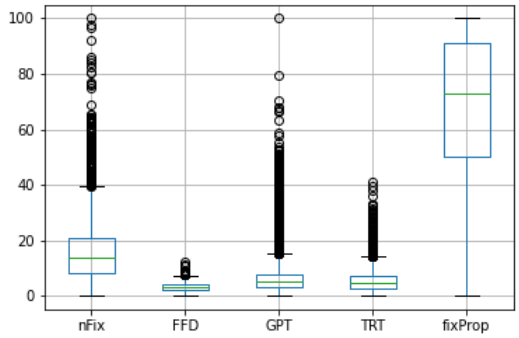

Figure 2: Boxplot showing the feature value distributions of the ZuCo dataset used in this paper.

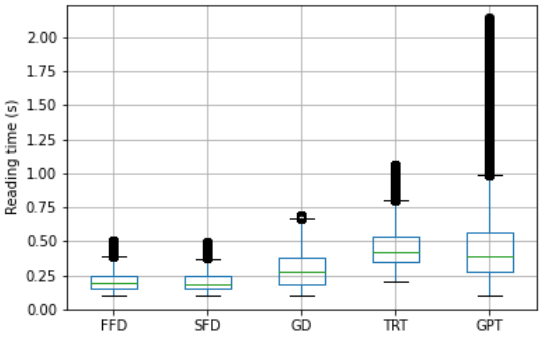

Figure 3: Boxplot of reading time data for English monolinguals of GECO.

fixProp (fixation proportion).

3.3 Ghent EyeTracking Corpus (GECO)

Ghent EyeTracking Corpus (GECO) (Cop et al., 2017) is a freely available English monolingual and Dutch–English bilingual eye-tracking dataset.

This paper only leverages the monolingual reading eye-tracking data which was recorded from 14 participants (6 males, 7 females, 1 other) which all are normal vision English monolingual undergraduates in Psychology from the University of Southampton, and none report any language or reading impairments. The average age of participants is 21.8 between 18 and 36. The data were recorded using the EyeLink 1000 system while participants read the English version of the novel *The Mysterious Affair at Styles*.

The monolingual GECO includes 6,633 unique words, 5 eye-tracking features from it were used: FFD, SFD (single fixation duration), GD (gaze duration), TRT, GPT. The feature value distribution is shown in Figure 3. (Cop et al., 2017) did not report information about ethics approval.

4 Methodology

This paper implemented a bidirectional LSTM model with Bahdanau attention, 3 layers of 500 hidden units, and no drop out. All experiments are based on this architecture, but different word embedding approaches are used.

4.1 Baseline

Baseline model was trained using the word embeddings which initialized with pre-trained word type-level vectors in dimension 300 trained using fastText (Bojanowski et al., 2017).

4.2 fastText with ZuCo

The word embeddings were initialized with both fastText pre-trained word vectors the same as the baseline and the 5 eye-tracking features on word-level provided in the ZuCo dataset. Since the same word in different sentences has the different gaze feature values, in order to obtain a word-level value, I first average the different values of a same word as its final value. Then the 5 word-level gaze features were normalized and then appended to the fastText word vectors to construct the word embedding matrix of 305 dimensions for training.

This paper employs two methods to deal with the out-of-vocabulary (OOV) problem: When encountering words that do not appear in the eye-tracking data, (a) using the average value of each gaze feature to fill, or (b) using the zero-vector to fill.

4.3 fastText with GECO

This approach of combining fastText vectors with GECO eye-tracking features is the same as the approach described in Section 4.2.

4.4 fastText with ZuCo and GECO

Eye-tracking features from ZuCo and GECO with fastText word vectors were used simultaneously to train the model. The 10 word-level gaze features were normalized and then appended to the fastText word vectors to construct the word embedding matrix of 310 dimensions for training, and used the zero-vector filling method for OOV.

5 Experiments

This paper split the Amazon Fine Food Reviews dataset of 393,222 data into 70% training data, 20% test data, and 10% validation data.

5.1 Evaluation Metrics

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a common evaluation method for automatic text summarization task proposed by (Lin, 2004). This measure counts the number of overlapping units such as n-gram between the gold summary and the candidate summary. This paper used the F_1 score of ROUGE-1, ROUGE-2, and ROUGE-L for evaluation.

5.2 Implementation Detail

This paper implemented and trained the models under the Keras architecture. All models are trained for 50 epochs on a Tesla-P100-16G GPU with batch sizes of 512 and used Early Stopping to halt the training process. The optimizer is RMSprop with a learning rate of 0.001. All experiments kept the setup and hyperparameters the same with the only difference of word embedding approaches.

5.3 Experimental Results

The experimental results are shown in Table 1. The first section contains the baseline model and the second section contains the models with different word embeddings. From Table 1 we can see that the models with gaze feature all significantly outperform the baseline model. Filling the OOV word vectors of eye-tracking data using zero vectors is better than using average value. The word embedding that combines fastText, Zuco, and GECO achieves the best performance, and the baseline model without gaze feature performed worst. There is no significant difference between ZuCo and GECO on the model. These results suggest that zero vector filling method is more suitable for eye-tracking data, and gaze features can improve text generation task.

6 Discussion

6.1 Difference between Results

Although the experimental results already showed that the eye-tracking data significantly improved the baseline model's performance, this paper also concerned about whether there are statistically significant differences within the results. I proposed the following hypotheses, H_0 : There is no difference between the means of the results of different models; H_1 : The means of results of the models are not all equal.

I applied ANOVA (Analysis of Variance) (Miller Jr, 1997) on the experimental results, and the analysis results are shown in Table 2. For fastText-ZuCo and fastText-ZuCo+GECO, we can reject the null hypothesis at a 95% level of confidence, which means that there are significant differences between the results of the models. The word embeddings with gaze features from ZuCo made a significant difference with the baseline model. But for other model pairs, there is no significant difference between the results even though they got different Rouge scores.

Model1	Model2	p-value	reject
fastText	ZuCo	0.011	True
fastText	GECO	0.285	False
fastText	ZuCo+GECO	0.010	True
ZuCo	GECO	0.140	False
ZuCo	ZuCo+GECO	0.982	False
GECO	ZuCo+GECO	0.135	False

Table 2: The results of ANOVA on the different groups of the models. A p value of less than 0.05 is required for significance and reject the null hypothesis.

6.2 Limitations

Compared with text annotation, eye-tracking data collection is more difficult and complex, which leads to a low resource problem. ZuCo and GECO are the relatively largest eye-tracking datasets in existence, but they still cover far fewer words than words in pre-trained word vectors using fastText. The eye-tracking dataset used in this paper comes from some different tasks than ATS. It performs well to do experiments directly on the original text of the eye-tracking datasets, but if transfer these gaze features to other tasks and datasets will face a serious OOV problem. Many words in NLP datasets are not appeared in eye-tracking dataset, so it's hard to take full advantage of gaze features for different NLP tasks.

Even though this paper effectively improved baseline model performance with eye-tracking data, I argue that eye-tracking data has the potential to provide much higher results than that since their current results are based on eye-tracking data with low word coverage. If an eye-tracking dataset containing enough words is available, it is reasonable to believe that there would be much better results.

For future work, a more efficient approach to transfer eye-tracking data to other tasks may solve the OOV problem and bring a greater improvement, which also helps gaze features can be more commonly used in NLP tasks.

Model	ROUGE-1	ROUGE-2	ROUGE-L
LSTM			
w/ fastText	13.316	2.985	13.151
w/ fastText + ZuCo (zero)	14.125	3.338	13.968
w/ fastText + GECO (zero)	14.126	**3.372**	13.967
w/ fastText + GECO (average)	13.936	2.991	13.779
w/ fastText + ZuCo (zero) + GECO (zero)	**14.336**	3.305	**14.161**

Table 1: F_1-score ROUGE on the test set. The '(zero)' stands for the OOV word vectors of eye-tracking feature are filled with zero vectors, the '(average)' stands for using average value filled the word vectors.

7 Conclusion

This paper implemented a bi-LSTM model with different types of word embeddings for abstractive ATS and presented the results. The experimental results show that the models with eye-tracking data all outperform the baseline model, and the model with ZuCo and GECO data achieved the best performance. Then an ANOVA test was applied and the analysis results showed that there is a statistically significant difference between the results of the baseline and the model with ZuCo. Gaze features significantly improved the text generation performance on abstractive Automatic Text Summarization.

References

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2):602–615.

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 Human Language Technology Conference of the NAACL, Main Conference*, pages 296–303.

Matthew Green. 2014. An eye-tracking evaluation of some parser complexity metrics. Association for Computational Linguistics.

Michael Hahn and Frank Keller. 2016. Modeling human reading with neural attention. *arXiv preprint arXiv:1608.05604*.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019a. Advancing nlp with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*.

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*, 5(1):1–13.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2019b. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. *arXiv preprint arXiv:1912.00903*.

Sigrid Klerke, Héctor Martínez Alonso, and Anders Søgaard. 2015. 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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908.

Rupert G Miller Jr. 1997. *Beyond ANOVA: basics of applied statistics*. CRC press.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y 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.

Kun Yi, Yu Guo, Weifeng Jiang, Zhi Wang, and Lifeng Sun. 2020. A dataset for exploring gaze behaviors in text summarization. In *Proceedings of the 11th ACM Multimedia Systems Conference*, pages 243–248.