A Dataset for Exploring Gaze Behaviors in Text Summarization

Kun Yi, Yu Guo, Weifeng Jiang, Zhi Wang
Tsinghua Shenzhen International Graduate School
{yik17@mails., guoy18@mails., jwf18@mails.,
wangzhi@sz.}tsinghua.edu.cn

Lifeng Sun
Tsinghua University
sunlf@tsinghua.edu.cn

ABSTRACT

Automatic text summarization has been a hot research topic for years. Though most of the existing studies only use the content itself to generate the summaries, researchers believe that an individual's *reading behaviors* have much to do with the summaries s/he generates, usually regarded as the ground truth. However, such research is limited by the lack of a dataset that provides the connection between people's reading behaviors and the summaries provided by them. This paper fills in this gap by providing a dataset covering 50 individuals' gaze behaviors collected by a high-accurate eye tracking device (that generates 100 gaze points per second) when they are reading 100 articles (from 10 popular categories) and composing the corresponding summaries for each article. Collected in a controlled environment, our dataset with 157 million gaze points in total, provides not only the basic gaze behaviors when different people read an article and compose its corresponding summary, but also the connections between different behavior patterns and the summaries they will provide. We believe such a dataset will be valuable for a wide range of studies, and we also provide sample use cases of the dataset.

CCS CONCEPTS

• **General and reference** → *Evaluation*; • **Information systems** → **Users and interactive retrieval**.

KEYWORDS

Dataset, Text Summarization, Personalized Text Summarization, Gaze Behaviors

ACM Reference Format:
Kun Yi, Yu Guo, Weifeng Jiang, Zhi Wang and Lifeng Sun. 2020. A Dataset for Exploring Gaze Behaviors in Text Summarization. In *11th ACM Multimedia Systems Conference (MMSys'20), June 8–11, 2020, Istanbul, Turkey*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3339825.3394928

1 INTRODUCTION

Automatic text summarization has been a hot research topic for years. Thanks to the continuous emergence of various datasets, the performance of automatic text summarization models has been improved significantly. Though most of the existing studies only use the content itself to generate the summaries, it is believed that an individual's *reading behaviors* have much to do with summaries s/he generates, usually regarded as the ground truth. Nowadays, with the continuous innovation of eye tracker devices, we can easily and accurately obtain people's gaze points. And there have been some studies on other tasks of natural language processing, which introduce people's reading behaviors to improve the design of models and achieve performance improvements. For example, a two-stage reading behavior model [14] significantly improved performance in the machine reading comprehension task.

However, because of the lack of a dataset that provides the connection between people's reading behaviors and the summaries provided by them, there was no such research in the automatic text summarization task. The existing datasets, such as CNN / Daily Mail, New York Times, only contain the article content and the corresponding summary.

This paper fills in this gap by providing a dataset covering 50 individuals' gaze behaviors when they are reading 100 articles from 10 categories. We used three high-accurate eye tracking devices, which can generate 100 gaze points per second to record participants' gaze behaviors. The entire data collection experiment is performed in a controlled environment. After data cleaning and calibration, our dataset with 157 million gaze points in total, provides not only the basic gaze behaviors when different people read an article and compose its corresponding summary, but also the connections between different behavior patterns and the summaries they will provide. Based on the dataset, we made some basic measurements and observations, and provide some use cases for it:

- Our measurements show that users have their own stable reading patterns. We can try to integrate reading patterns into the design of automatic text summarization models.
- Our measurements also show that there are differences in reading patterns among different people, as well as among the corresponding summaries. We can try to dig out the correlation between the two differences, which will be helpful in improving the performance of personalized automatic text summarization.

The rest of the paper is organized as follows. First, a brief overview of related works is presented in Section 2. Then, Section 3 describes the detailed process of data collection. Section 4 gives an overview of the dataset. Section 5 presents the availability and format of the dataset. Section 6 provides some basic measurements based on the dataset. Finally, some conclusions are provided in Section 7.

2 RELATED WORK

Recent studies have made great progress in text summarization, especially abstractive text summarization methods. Rush et al. [11] and Nallapati et al. [8] applied abstractive text summarization models based on RNN and attention mechanism. See et al. [12] proposed

MMSys'20, June 8–11, 2020, Istanbul, Turkey

ACM ISBN 978-1-4503-6845-2/20/06…$15.00
https://doi.org/10.1145/3339825.3394928

a pointer generator model with coverage to solve the problem of out-of-vocabulary (OOV) words and duplicate words. Chen et al. [1] used reinforcement learning to generate a concise overall summary. With the success of pre-trained language models in various natural language processing tasks, some studies (Liu et al. [5], Zhang et al. [13]) proposed models based on pre-trained language models to generate fluent and informative summaries.

Nowadays, a few studies have begun to apply gaze behaviors to various tasks of natural language processing, and have achieved a good improvement. Some of them use gaze behaviors to evaluate the readability [6] and grammar [4] of text. Mishra et al. [6] proposed to predict the rating of text quality using gaze behaviors. Mishra et al. [7] improved the performance of the model in cognition-cognizant sentiment analysis by introducing an auxiliary task of predicting the gaze time on words. Zheng et al. [14] thoroughly investigated human behavior patterns during reading comprehension tasks and proposed a two-stage reading behavior model, which significantly improved performance in the MRC task. However, as far as we know, no study has been done to introduce gaze behaviors into the text summarization task.

Several datasets have been proposed for text summarization, such as Gigaword [9], XSum [10], CNN / Daily Mail Dataset [3]. But these datasets only have articles and corresponding summaries, and they do not contain data that can be used to study the process of reading and summarizing articles in detail. To address this deficiency, we used several eye tracking devices to collect users' gaze behaviors during text summarization and present a new dataset.

3 DATASET COLLECTION

This section mainly introduces the process of data collection. The whole procedure of data collection is shown in Fig. 1.

3.1 Resource preparation

We manually collected a total of 100 samples in 10 categories from the public news websites[1] for gaze collection. And Each sample includes an article and a title, where the title is used as a reference when the user writes the summary. More specific usage details will be explained in Section 4.

We used three eye tracker devices[2] to capture the gaze behavior of participants when they were reading articles. People usually use it in PC gaming for an enhanced streaming, gameplay and esports experience. We used an open-source program to obtain the original acquisition data from the eye tracker. And the data sampling frequency is about 100Hz.

3.2 Data Collection Procedure

The entire data collection process is the same for each participant and is divided into four stages:

- **Equipment Calibration:** Before starting to read articles, every participant needs to complete the calibration of the device according to the instructions of the eye tracker. For each participant, this step only needs to be performed once before the experiment begins.
- **Document Reading:** After the device calibration is completed, the participant enters the article reading stage. At this stage, participants are required to read the entire article as continuously as possible, with no limit on reading time. We have written a webpage program to display articles. By adjusting the font size and the spacing between Chinese characters, the gaze data returned by the eye tracker can be accurate to the level of words. In addition, when displaying the content of the article, we put all the Chinese characters of the same word on the same line. This avoids the jump in reading sight caused by a word appearing in two lines.
- **Document Summarizing:** After reading the article, the participant is asked to give a summary that s/he thinks fits the content of the current article. The participant can use words that do not appear in the original text to compose the final summary. At this stage, the original title of the article is presented as a reference summary at the bottom of the screen. The reason for providing a reference summary is to prevent the summary given by the participant from being too ad-hoc.
- **Data Cleaning and Calibration:** After the current user has read and summarized all the articles, we can get eye-tracking logs of the entire reading process through the background program. First, we used timestamps recorded by the webpage program to obtain the log corresponding to each article. In actual experiments, it is found that the eye tracker still has a certain deviation from the text near the edge of the screen, which requires us to perform data calibration based on the original data. We just need to get the coordinates returned by the eye tracker on the four corners of the reading frame and calculate the offset. And a heat map of the gaze behaviors on the article is shown in the upper right corner of Fig. 1.

We divided the experiment into two sub-experiments, each of which included 50 articles. In order to avoid fatigue affecting data quality, each participant was required to have two sub-experiments at two different times. Articles used in each sub-experiment are out of order and shuffled every time.

4 DATASET OVERVIEW

4.1 Articles Used in the Study

Articles used in the study were collected manually from public news websites. There are 100 articles in total which belong to 10 popular categories. Each category has 10 articles. When selecting articles, we deliberately avoided articles that can summarize the entire article content in the first sentence. In addition, articles that are too short or too long were not selected. The average length of all articles is around 502 Chinese characters, and that of all titles is around 22 Chinese characters. The longest article has 842 Chinese characters, and the shortest article has 99 Chinese characters. Fig. 2 shows the length distribution of articles in different categories.

4.2 Participants

We recruited a total of 50 users, and Table 1 presents the demographic profile of all the participants. All the participants are postgraduates from various departments, of which 64% are female, and they are all around 23 years old.

[1] Netease News and Tencent News
[2] Tobii EyeTracking 4C

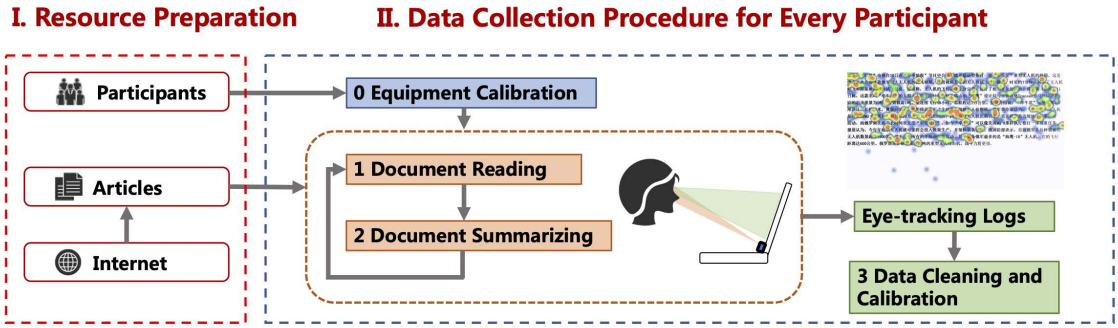


Figure 1: The procedure of data collection

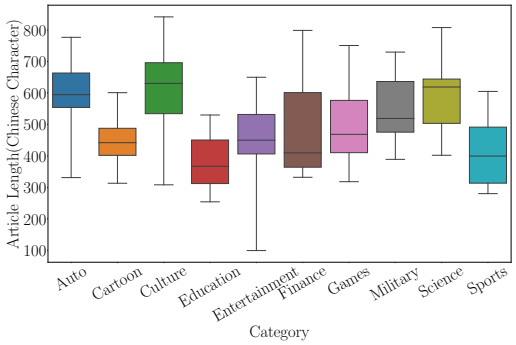


Figure 2: The article length distribution in different categories.

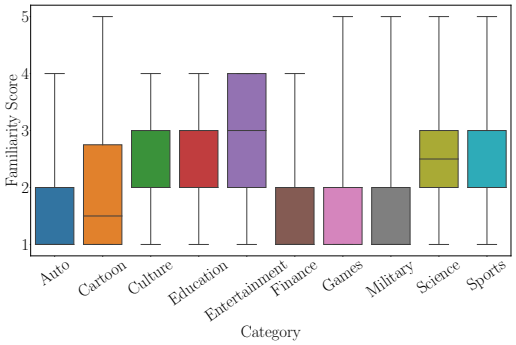


Figure 3: The familiarity distribution in different categories.

Gender		Age				
Male	Female	21	22	23	24	25
18	32	1	16	17	9	7

Table 1: Demographic Profile of the Participants

In addition, all participants were asked to fill out a questionnaire to count participants' familiarity with different categories of articles. The familiarity score ranges from 1 to 5, where 1 means very unfamiliar, and 5 means very familiar. Fig. 3 shows the distribution of all participants' familiarity with different categories of articles. It can be seen that the familiarity for most categories is below 3. For articles of unfamiliar categories, the user does not have much background knowledge to refer to when giving the corresponding summaries. So the user need to spend more time reading articles and understanding content.

4.3 Gaze Data Collected

After data cleaning and calibration, we obtained a total of 157 million gaze points, which is about 437 hours. Because the participants were required to complete the reading of the article in accordance with their daily reading habits in the experiment, there was a large difference in the gaze time of each participant. Fig. 4 shows the average gaze time on each Chinese character of each participant and Fig. 5 shows the average gaze time on each Chinese character in different categories.

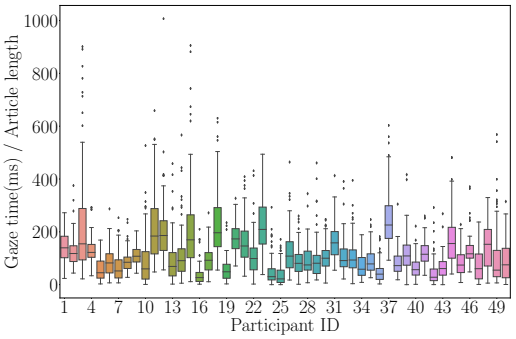


Figure 4: The average gaze time on each Chinese character of each participant.

4.4 Summaries Given by Participants

After reading the article, participants were asked to give their summary of the article. The overall average length of the final summaries is about 21 Chinese characters, which is similar to the average length of the titles of 22 Chinese characters. The former has a wider range of lengths, the shortest is only four Chinese characters, and the longest is 85 Chinese characters. But the latter has a minimum of 10 Chinese characters and a maximum of 31 Chinese characters. It means that artificially given summaries are more diverse and challenging.

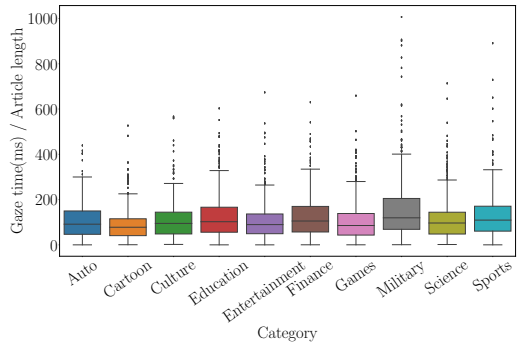


Figure 5: The average gaze time on each Chinese character in different categories.

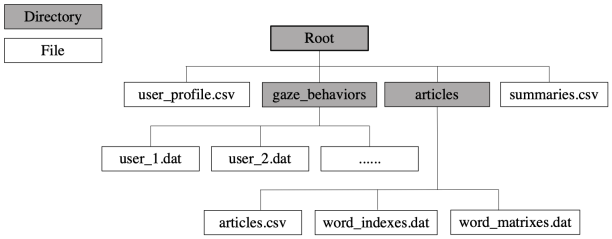


Figure 6: The Directory Structure of the Dataset

Data	Articles	Participants	Gaze Records	Summaries
Nums	100	50	4791	4998

Table 2: Statistics of our dataset

5 DATASET AVAILABILITY AND FORMAT

The dataset can be downloaded from our repository[3]. The size of our dataset is about 191 MB after compression. Fig. 6 shows the hierarchical structure of our dataset. The dataset mainly includes four parts, which are user profiles, articles used, effective gaze behaviors after preprocessing, and summaries given by participants. The specific format of the files contained in the dataset is explained below. Table 2 shows the statistics of our dataset. It should be noted that files with a *.dat* extension are binary files. You need to load data from them using the python module *pickle*. And the *pickle* module can transform a complex object into a byte stream.

5.1 Articles Used

5.1.1 articles.csv.

This file stores the details of the 100 articles used in the experiment. Every three lines constitute a unit, where the first line is the Article_ID, the second line is the title of the article, and the third line is the content of the article.

5.1.2 word_indexes.dat.

We used the Chinese word segmentation tool to segment the article and saved the results in this file. The format of the data object loaded by pickle is as follows:

```
dict(
  Article_ID : dict(
    Word_index : tuple(Word, POS)
  )
)
```

- **Word_index**: refers to the word index after segmenting each article.
- **POS**: refers to the part of speech corresponding to the word.

5.1.3 word_matrixes.dat. :

- As shown in the second stage in Fig. 1, we displayed the content of the article on a computer monitor. In the experiment, in order to map the gaze behaviors to words in the later stage, we divided the display area into regions of 102 columns by 18 rows, and each unit contained a standard Chinese character.
- Then, each Chinese character was displayed in the corresponding unit in order. Because Chinese words generally include multiple Chinese characters, if all Chinese characters for a word were not displayed on the same line, we chose to display the word as a whole on the next line, which ensured that the participants' reading was more consistent.
- The *word_matrixes.dat* stores the layout of each article content on the display. The specific format of the data object loaded by pickle is as follows:

```
dict(
  Article_ID : Layout_array()
)
```

The **Layout_array** is a Numpy array of 102 columns by 18rows. Each value of the array represents the index of the word to which the Chinese character at the current position belongs. And when the value is -1, it means that the Chinese character at the current position is a blank.

5.2 Participants' Profiles

The profiles of all participants are stored in the *user_profile.csv* file in csv format(13 fields):

- **User_ID**: Unique ID for each participant, from 1 to 50.
- **Age, Gender**: Participant's age (integer format) and gender.
- **Military, Entertainment, Sports, Science, Culture, Finance, Cartoon, Games, Auto, Education**: The familiarity scores with 10 different categories of articles. Each score ranges from 1 to 5 in the integer format.

5.3 Gaze Behaviors

The gaze behaviors of different participants are stored in different files after being pre-processed. Each file is named *user_{User_ID}.dat* according to the participant's ID. The specific format of the data object loaded by pickle is as follows:

```
dict(
  Article_ID : list[
    tuple(Gaze_index, Word_index, Array_index)
  ]
)
```

[3]https://github.com/MMLabTHUSZ/ADEGBTS

- **Gaze_index**: refers to the index of the gaze point collected by the eye tracker. Because the gaze points beyond the article display area have been filtered out in the preprocessing stage, the Gaze_index list may be discontinuous and not start with 1.
- **Word_index**: refers to the index of the word to which the current gaze point belongs.
- **Array_index**: refers to the position of the current gaze point. If the current position is in the h-th column of the w-th row, the corresponding Array_index is equal to $w * 102 + h$.

We regarded gaze behaviors of a participant on an article as a record. After preprocessing, a total of 4791 valid records were obtained.

5.4 Summaries Given by Participants

All summaries given by participants are stored in the *summaries.csv* in csv format(3 fields):

- **Article_ID**: Unique ID for each article, from 00 to 99.
- **User_ID**: Unique ID for each participant, from 1 to 50.
- **Summary**: The corresponding summary given by the participant.

A total of 4998 valid summaries were collected in the experiment.

6 DATASET VISUALIZATION AND MEASUREMENT

6.1 Gaze Distribution

In order to compare the similarities and differences in the gaze distribution of different people during reading, we show the collected gaze behaviors in the form of a heat map. The brighter part of the heat map indicates that the participant has been reading the current area for a longer time. Fig. 7, Fig. 8, and FIg. 9 show the gaze distributions of three participants when they completed the text summary task of three articles, respectively. The three articles shown in the figures belong to different categories. Article 00 in Fig. 7 belongs to the military category, article 77 in Fig. 8 belongs to the games category, and article 92 in Fig. 9 belongs to the culture category.

Fig. 7 shows the gaze distribution of 3 participants when they were reading article 00. It can be seen from the results that in the process of reading the same article, different participants have different emphasis on the article. Participant 10 was more inclined to read the content at the end of the article, Participant 37 was more inclined to read the content at the beginning and end of the article, and Participant 8 had no obvious reading tendency. From the results in Fig. 8 and Fig. 9, we can also see the different focus of different participants.

When we look at these heat maps from the perspective of participants, it is not difficult to find that the same person may have a certain reading mode when reading different articles. It can be seen from Fig. 7 to Fig.9 that participant 8 prefers to read through the entire article, participant 10 prefers to read the beginning or end of the article, and participant 37 prefers to read both the beginning and end of the article together.

By comparing these groups of heat maps in the figures, we can propose the following two assumptions:

- When reading and summarizing text, everyone has their own stable reading patterns and preferences.
- When reading and summarizing text, there are different reading patterns and preferences existing between different people.

Fully mining the same points and differences between user reading patterns is more conducive to the design and improvement of automatic text summarization models and personalized text summarization models.

6.2 Diversity of Summaries

We already mentioned the length distribution of the summaries given by the participants in Section 4. In order to better compare the abstracts given by different participants from the semantic aspect, we fine-tuned a Chinese sentence similarity calculation model based on BERT [2]. We first calculated the similarities between different summaries given by participants corresponding to the same article. Fig. 10 shows the summary similarity distribution in different categories. It is not difficult to see that there are large differences between the summaries, and many of the similarities are lower than the empirical value of 0.8. The distributions in different categories are also different, among which summaries in the cultural category have the lowest similarity. There may be a correlation between the differences between the summaries given by the users and the differences in their gaze behaviors.

7 CONCLUSION

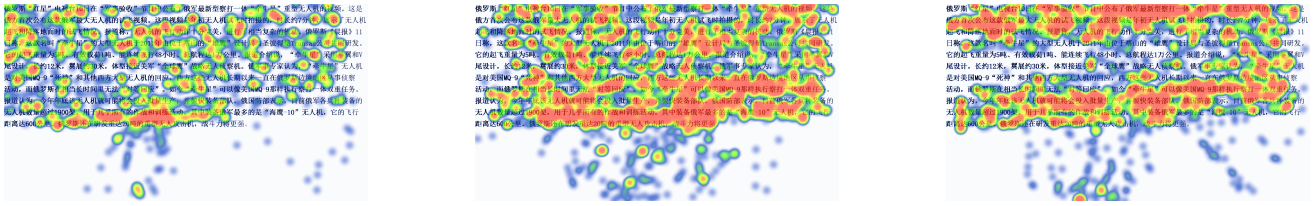
In order to facilitate the development of automatic text summarization, we present a gaze behavior dataset of user reading and summarizing articles. We also briefly explore the differences between the gaze behaviors of different participants and the differences between the given summaries. This dataset can be used to better explore the design of text summarization models and personalized text summarization models.

ACKNOWLEDGEMENT

This work is supported in part by NSFC under Grant No.61872215, SZSTI under Grant No.JCYJ20180306174057899, and Shenzhen Nanshan District Ling-Hang Team Grant under No. LHTD20170005. We would also like to thank InfleXion Lab for sponsoring this research.

REFERENCES

[1] Yen-Chun Chen and Mohit Bansal. 2018. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 675–686.
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.
[3] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*. 1693–1701.
[4] 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)*. 97–105.
[5] Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3721–3731.

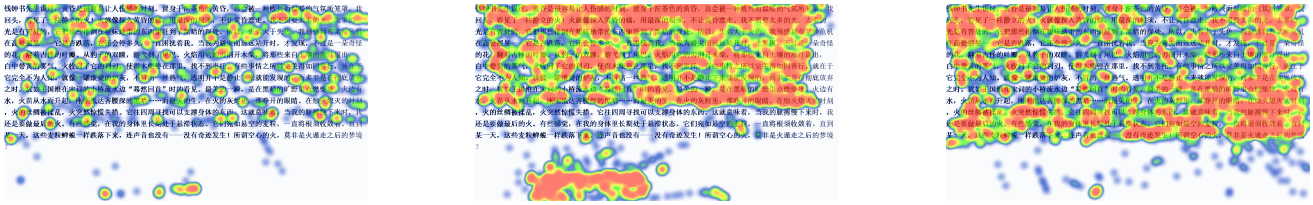


(a) Participant 8 (b) Participant 10 (c) Participant 37

Figure 7: Participants' gaze distribution when reading article 00

(a) Participant 8 (b) Participant 10 (c) Participant 37

Figure 8: Participants' gaze distribution when reading article 77



(a) Participant 8 (b) Participant 10 (c) Participant 37

Figure 9: Participants' gaze distribution when reading article 92

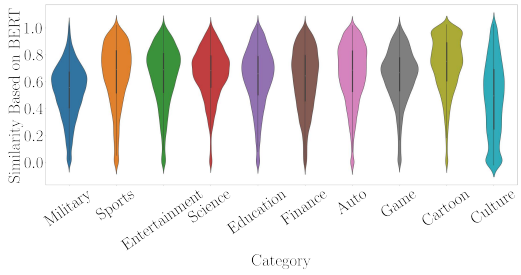


Figure 10: The summary similarity distributions in different categories.

[6] Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Scanpath complexity: Modeling reading effort using gaze information. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[7] 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 *Thirty-Second AAAI Conference on Artificial Intelligence*.

[8] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. 280–290.

[9] Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*. Association for Computational Linguistics, 95–100.

[10] Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1797–1807.

[11] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 379–389.

[12] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1073–1083.

[13] Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. 2019. Pretraining-Based Natural Language Generation for Text Summarization. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. 789–797.

[14] Yukun Zheng, Jiaxin Mao, Yiqun Liu, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Human behavior inspired machine reading comprehension. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 425–434.