論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Interpreting Reading and Writing Process of Neural Models using Eye-gaze Information
著者(和文)	Fariz Ikhwantri
Author(English)	Fariz Ikhwantri
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12915号, 授与年月日:2024年9月20日, 学位の種別:課程博士, 審査員:徳永 健伸,岡崎 直観,村田 剛志,齋藤 豪,井上 中順
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12915号, Conferred date:2024/9/20, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

TOKYO INSTITUTE OF TECHNOLOGY

DOCTORAL THESIS

Interpreting Reading and Writing Process of Neural Models using Eye-gaze Information

---

Author:
Fariz Ikhwantri

Supervisor:
Prof. Takenobu TOKUNAGA

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Engineering

in the

Department of Computer Science
School of Computing

August 21, 2024

Abstract

Interpretation of deep neural network models is an essential topic in the natural language processing (NLP) community. Yet, the relationship between models' and human behaviour in downstream tasks remains largely unexplored, calling for further investigation. Past studies have proposed various interpretation methods for neural networks, which provide saliency scores of input elements as clues to interpret a model's behaviour. On the other hand, eye movement research has a long and successful history of studying human cognitive processes. Although we cannot directly observe human cognitive processes, eye movement is believed to be a good proxy for reflecting them. Against such a background, it is natural to understand the neural network behaviour in solving NLP tasks by comparing it with the human eye-movement behaviour. This research investigates the alignment between saliency scores from neural network interpretation methods and human eye-gaze features from humans across diverse NLP tasks. Notably, this research extends beyond reading tasks like sentiment analysis, relation classification, and question answering to include writing tasks, such as summarisation.

This research aims to answer two research questions to understand the similarities and differences between models and humans in the decision process. The first question is, "Does the input word saliency from interpretation methods conform with human eye-gaze features?". The second question is, "How does the model saliency conformity impact model prediction?".

The first study is the task-specific reading. Four interpretation methods – simple gradient, integrated gradient, input-perturbation, and attention – were evaluated across three architectures: LSTM, CNN, and Transformer. Two publicly accessible corpora annotated with eye-gaze information, namely ZuCo and MQA-RC datasets, were utilised for this study. To answer the first question, I compared the models' input word saliency distance (SD) to human eye-gaze features. SD is defined as KL-divergence between the saliency distribution and eye-gaze feature distribution over input words. The results show that the Transformer has the highest similarity to the human gaze across reading tasks in most cases.

For the second question, I proposed a novel evaluation method called the "Saliency Distance-performance curve" (SDPC). This method visualises the cumulative model performance in relation to the SD scores. The SDPC sheds light on the underlying phenomena that were previously overlooked when solely relying on macroscopic metrics, such as average SD scores and rank correlations, as commonly done in past studies. Overall, the analysis of task-specific reading reveals that the impact of good saliency conformity between

humans and machines on task performance varies based on task combinations, interpretation methods, and architectures. These findings are crucial to consider when incorporating eye-gaze information for model training to enhance overall model performance.

In the writing task, I adopt summarisation as a target task because it involves reading a source text and writing its summary that captures the main ideas of the original text. Prior studies have analysed model interpretation in generation tasks such as translation or summarisation tasks. However, no study has addressed how the generation process compares to that of humans. The main challenge is aligning the model saliency output in the generation process and eye-gaze features from writing activity data. The model saliency output in the generation process is a matrix form, while eye-gaze features in reading or writing tasks are vectors, whose dimension corresponds to the words of the original text. I proposed a new framework for analysing summarisation models by comparing them to eye movement. The framework involves macroscopic and microscopic views of model saliency and human gaze data to handle the different representations. The model saliency output is transformed to the same representation of eye-gaze features, a vector, in the macroscopic analysis. On the other hand, eye-gaze features are transformed to the same representation of the model saliency, a matrix, in the microscopic analysis. In this study, I also built a novel dataset of extractive and abstractive summarisation by language learners, which is annotated with eye-gaze information and keystroke logs.

To answer the first question, I investigate the rank correlations between model saliency scores and human fixation counts in the macroscopic and microscopic analyses. Our findings suggest attention-based saliency scores partially align with human fixation counts.

For the second research question, I propose an ablation analysis which removes part of the input according to the model saliency and the eye-gaze feature. The macroscopic ablation analysis indicates that removing important words according to the human gaze can impact model performance. However, microscopic ablation analysis reveals that the human gaze does not impact model performance, which differs from macroscopic ablation. This discrepancy may be attributed to the forced decoding method, which might not accurately reflect the prediction scenario. The forced decoding uses previous ground truths, introducing bias to the decoder module. As a result, while the human gaze may influence models when using its output decoding, forced decoding can lead models to rely heavily on previous ground truths, affecting their performance.

Acknowledgements

I would like to express my heartfelt gratitude to the many individuals and institutions who have supported and guided me throughout my PhD journey. First and foremost, I am deeply thankful to my family for their unwavering support, encouragement, and understanding. Their love and belief in me have been my anchor and motivation throughout this challenging yet rewarding endeavor. I am also grateful to my friends and colleagues whose companionship, advice, and camaraderie made this journey enjoyable and memorable.

Special appreciation goes to my supervisor, Prof. Takenobu Tokunaga, for their exceptional mentorship, patience, and invaluable guidance. His expertise and constructive feedback have not only been instrumental in shaping this thesis but also in shaping my growth as a researcher. I also wish to thank Dr. Hiroaki Yamada for all his help and support throughout my PhD research.

I would like to extend my gratitude to the Japan Ministry of Education, Culture, Sports, Science and Technology through the MEXT Scholarship. Their financial support enabled me to pursue this research and achieve my academic goals.

I am thankful to my friends at Tokunaga-lab for their constructive feedback and shared insights, which significantly enriched the research presented in this thesis. I am also thankful to the faculty members and administrative staff of the Tokyo Institute of Technology for their assistance, resources, and facilities that supported my research endeavors. My sincere appreciation also goes to the participants who generously shared their time, which was critical to the empirical findings of this thesis.

I extend my deepest appreciation to all those who have contributed, directly or indirectly, to this thesis. Your support and encouragement have been indispensable in my journey toward completing my doctoral studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Interpretation of deep neural network models is an essential topic in the natural language processing (NLP) community (Belinkov, Gehrmann, and Pavlick, 2020; Doshi-Velez and Kim, 2017; Lipton, 2018). However, little is still known about the relationship between models' decision-making and the human cognitive process on the downstream tasks. Past studies have proposed various interpretation methods (Guan et al., 2019; Ribeiro, Singh, and Guestrin, 2016; Simonyan, Vedaldi, and Zisserman, 2013; Sundararajan, Taly, and Yan, 2017), which provide saliency scores of input elements as clues to interpret a model's behaviour.

Eye movement research has a long and successful history of studying human cognitive tasks, including reading, scene perception and visual search (Rayner, 1998, 2009). Eye movements reside between perception and cognition; they play a central role in the visual system, simultaneously related to human cognitive processes such as memories, expectations and goals. Particularly, eye movements during reading have been studied in depth. Although we can not directly observe human cognitive processes, eye movements are believed to be a good proxy for reflecting them (Richardson, Dale, and Spivey, 2007). The eye-mind assumption (Just and Carpenter, 1980) tells us "*... the eye remains fixated on a word as long as the word is being processed. So the time it takes to process a newly fixated word is directly indicated by the gaze duration.*". Against such a background, it is natural to understand neural network behaviour in solving NLP tasks by comparing it with human eye movement behaviour. Figure 1.1 shows the word saliency of humans and machines in solving the sentiment analysis task. The colour density indicates how much humans and machines look at the word in solving the task. The left example[1] shows good conformity of both saliency, but the right example[2] does not. The machine's answers were correct (left) and incorrect (right) accordingly. We can notice that the machine less attends to the adjectives "silly" and "hilarious" in the right example, which helps judge the sentiment polarity. This

[1]Machine saliency is extracted from Transformer with the gradient interpretation method, and human saliency is based on the first fixation duration (FFD).

[2]Machine saliency is extracted from Transformer with the input perturbation method, and human saliency is based on FFD.

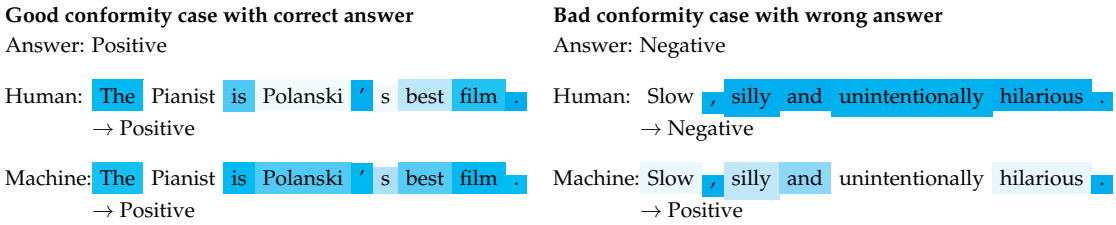


FIGURE 1.1: Example of word saliency of humans and machines in solving the sentiment analysis task

example shows that machines could improve their performance by learning to gaze at the same areas as humans.

There have been studies to investigate the interpretation methods in terms of human eye movement on reading activity. For instance, Hollenstein et al. (2019a), Sood et al. (2020b) and Hollenstein and Beinborn (2021a) investigated the correlation between saliency from the interpretation methods and eye movement data in the context of NLP tasks. Existing studies for the question-answering (QA) task suggested that models do not always reflect human eye movement behaviour despite showing good prediction performance (Feng et al., 2018; Sood et al., 2020b). However, these studies employed different datasets, interpretation methods and evaluation measures.

In contrast, eye movement studies on the writing process have been less studied in NLP fields. There are still large gap between understanding NLP models related to writing activity. Carl and Kay (2012) studied eye movement on translation process on human factors. A recent study (Sahoo and Carl, 2019) provides the monolingual data such as paraphrasing and summarization task, but did not focus on text generation models in NLP.

Recent progress on pre-trained Transformer-based summarisation models (Lewis et al., 2020; Liu and Lapata, 2019; Stiennon et al., 2020) improved the model performance by a large margin. Xu and Durrett (2021) investigated the inner-working process of a pre-trained transformer-based summarisation model by breaking it down into different parts of models, the pre-trained and finetuned stages, and the Transformer components. Their analysis shows the model bias towards pre-training data and memorisation. They also addressed the difference in representation for interpreting the classification and autoregressive generation model. However, they did not consider comparing the model and human behaviour.

In order to fill in the gaps and expand the empirical findings from the series of previous studies, I conduct experiments systematically combining three factors: NLP tasks, interpretation methods and neural network architectures for reading and writing activity. Specifically, the aim of this thesis is to answer the following research questions (RQ).

RQ1: Does the input word saliency from interpretation methods conform with human eye-gaze features?

RQ2: How does the model saliency conformity impact model prediction?

RQ1 concerns whether the machine looks at the same input elements as the human to solve the task. This question is interesting from an engineering viewpoint. If machines behave differently from humans and do not achieve human-level performance, machines have something to learn from human behaviour. RQ1 leads to RQ2, which concerns whether the machine which behaves like humans performs the task better. This question is interesting from both scientific and engineering viewpoints. From a scientific viewpoint, the comparison leads to an investigation of the human brain mechanism using neural networks as an operational and computational tool.

In this thesis, I analyse downstream NLP model interpretation methods with eye gaze as cognitive proxy. I focused on reading and writing tasks. There are two main reasons. First, reading and writing are essential activities in human cognition and languages. Other activities, such as visual search, speaking, or listening, can also be linked with eye movement, but they involve images and speech modality. This study concerns text modality in reading and writing activities. Second, the study interests is in language processing. I can control task complexity when considering various kinds of reading and writing tasks. Controlling task complexity aligns with the recent trends in eye-movement research towards task-specific reading (Hahn and Keller, 2018; Hollenstein et al., 2019a) from normal reading. This study also looks further into writing activity as the natural progression of task-specific and normal reading.

Summarisation is ideal for this study as it involves reading to distil important information and writing it down. Other writing tasks than summarisation include free-text writing and translation. Unlike translation, which involves multiple languages and has similar input and output lengths, summarisation typically results in shorter output lengths.

This thesis divided into two parts (i) Interpreting models in reading tasks and (ii) Interpreting models in summarization task.

In chapter 3, I provide the first broad overview of the relation between different interpretation methods and human eye movement behaviour across different architectures and tasks. I analyse four interpretation methods based on gradient (Ancona et al., 2018), integrated gradient (Shrikumar, Greenside, and Kundaje, 2017), input perturbation (Ribeiro, Singh, and Guestrin, 2016) and attention (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019), using common neural network architectures: Long-Short-Term Memory Network (LSTM) (Hochreiter and Schmidhuber, 1997), Convolutional Neural Network (CNN) (Kim, 2014) and Transformer (Vaswani et al., 2017). In the study, I compare the interpretation methods against human eye movement across three types of NLP tasks: sentiment analysis, relation classification, and question answering in reading tasks through our experiments. These findings have direct implications for the development of more efficient and accurate NLP models based on human cognition.

In chapter 4, I conducted a novel study comparing transformer summarisation models and human with eye-gaze information. The eye-gaze information was collected during a human summarisation activity, providing a unique perspective on the performance of these models. However, there are

large gap between model text generation and eye-movement in writing activity. To this end, I propose a framework to analyse neural summarisation models through comparison with human summarisation behaviour. I utilise human eye gaze and keystroke data as a proxy of human behaviour.

1.2 Contributions

This study does not aim to improve the model performance. This research proposes a framework to analyse models' behaviour by comparing with humans' eye movement. The contribution can be categorised into technical contributions and empirical findings.

1.2.1 Technical Contributions

The technical contributions of this thesis are as follows:

1. This study reproduce (Sood et al., 2020b) and integrate previous studies models and saliency evaluation to the same evaluation metric (Eberle et al., 2022; Hollenstein and Beinborn, 2021a; Sood et al., 2020b) and implementation in reading tasks. I defined the evaluation metric as Saliency Distance (SD) to measure conformity (RQ1) between model saliency and human gaze. SD can be implemented with KL-divergence to measure relative entropy or ranking correlation to measure relative similarity between the saliency distribution over input words and an eye-gaze feature. The usage depends on the settings and goal of evaluation. This standardise methods and procedures, making it easier to compare results and methodologies across different studies with different architectures and tasks.

2. I proposed a novel evaluation technique called "Saliency Distance(SD)-performance curve" (SDPC), which presents the cumulative model performance against the SD scores (Section 3.7.1). This method is proposed to provide more practical and simple approach to answer RQ2 in reading tasks. This method finds underlying phenomena that were overlooked using only macroscopic metrics such as rank correlation compared to related studies. This method can be used to improve machine performance by training them to behave like humans. This metric also can be used without re-running the model prediction from saliency output.

3. I proposed an evaluation framework to analyse the conformity between humans and machines and its relation to prediction performance using different decoding approaches in summarisation. The framework can be divided into macroscopic and microscopic settings for model saliency and human gaze. In macroscopic, both model and human gaze are represented as vectors similar with reading tasks approach. In microscopic, the analysis concern the generation process of a summarisation model where the saliency output is a sequence of vectors or

a matrix form. To answer RQ2, I used input ablation approach to both macroscopic and microscopic.

4. I proposed a feature representation to align model saliency and eye-gaze information to align with sequence output of model generation. I introducing the temporal-based segmentation for the time series of fixations from raw eye-gaze data (Section 4.7). These features can help in comparing the attention of humans and machines in other writing tasks such as translation.

5. I present a collection of 53 summaries with eye-gaze information from 30 participants in extractive and abstractive summarisation tasks (Section 4.3.3). The analysis shows some intriguing behaviours of language learners that are different from native speakers in summarisation tasks. This dataset expands the possibility of analysing eye gaze in L2 writing tasks.

1.2.2 Empirical Findings

This study provides a thorough analysis of NLP model interpretability with eye movement in reading tasks (Chapter 3). I provide the first broad overview of the relation between different interpretation methods and human eye movement behaviour across different architectures and tasks. This is the first study to analyse the interpretability of model behaviour in summarisation by comparing it to human behaviour in the same task using eye-gaze data (Chapter 4). From the research questions. The general answer for tendencies of this study findings as follows:

RQ1 Reading: I found that Transformer models tend to have the highest conformity with human attention across tasks in most cases. The attention as interpretation method also has the highest combination with Transformer or other models such as CNN.

RQ1 Summarisation: I found some positive conformity to the human gaze. The attention as interpretation method also has the highest combination with either pretrained or finetuned version of the models.

RQ2 Reading: In reading tasks, model saliency conformity with human gaze does not always lead to positive relation with the model performance. In relation classification task, Gradient-based method interpretation shows positive relation with LSTM, CNN and Transformer architectures.

RQ2 Summarisation: In summarisation, I found the model's performance can be affected by the human gaze in macroscopic ablation. However, microscopic ablation analysis demonstrates negative results due to bias of previous ground-truth input to the model when following previous study (Xu and Durrett, 2021)

This study's findings underscore the variability and intricate interplay between tasks, interpretation methods, and architectures in leveraging eye-gaze information for model training, providing critical insights for developing robust and efficient NLP models to achieve better performance through human-like cognitive mechanisms.

Chapter 2

Related Work

2.1 Eye-movement as Cognitive Proxy

Eye movement has been considered as a proxy for cognitive activity load (Rayner, 1998). Eye movement is obtained by the eye-tracking device that records a series of time-stamped **eye-gaze** coordinates on the screen. When a series of gaze points are close in time and space, such gaze point clusters are called *fixations*. The eye-movement data is a time series of fixations from which various eye-gaze features can be derived by considering each input word as an area of interest (AOI).

These eye-gaze features derived from reading provide valuable insights into cognitive processes. Analysis of fixation duration, saccade length, and pupil dilation is commonly used to understand comprehension levels, cognitive load, and engagement. Previous studies have shown that longer fixation durations are associated with higher cognitive load and information processing difficulty, while shorter fixations and rapid saccades indicate fluid reading and understanding (Just and Carpenter, 1980).

2.1.1 Eye-gaze in Reading activity

Past researches on eye movement as cognitive proxy are focused on analyzing the linguistic effects such as ambiguity (Altmann, Garnham, and Dennis, 1992; Frazier and Rayner, 1982) and on collecting data on normal reading activity (Kennedy, 2003). It was found that word-level gaze features were related to linguistic and lexical characteristics of a word, such as word frequency and length (Just and Carpenter, 1980), as well as syntax and morphology (Frazier and Rayner, 1982; Hyönä, Bertram, and Pollatsek, 2004). Recently, there have been two main trends of eye-movement corpora. One is studying eye-gaze in multilingual reading activity (Berzak et al., 2022; Cop et al., 2017; Siegelman et al., 2022). Another one is task-specific reading such as sentiment detection (Hollenstein et al., 2019b), sarcasm (Mishra, Kanojia, and Bhattacharyya, 2016), question answering (Malmaud, Levy, and Berzak, 2020; Sood et al., 2020b). This research focuses on task-specific reading because it is related to the use of eye-gaze in the NLP model for downstream tasks.

2.1.2 Eye-gaze in NLP

Recent studies found that eye gaze focuses on text parts with different linguistic properties across tasks (Barrett et al., 2018a; Hahn and Keller, 2018). These findings made eye-gaze data from reading useful to enhance model performance and reduce the need for extensive training data in NLP. Other studies have demonstrated that eye-gaze features can help improve model prediction in downstream tasks such as co-reference resolution (Cheri, Mishra, and Bhattacharyya, 2016) and named entity recognition (Hollenstein and Zhang, 2019; Tokunaga, Nishikawa, and Iwakura, 2017) in sequence labelling tasks. Another study shows that eye gaze related to predicate-argument structure in a sentence (Maki, Nishikawa, and Tokunaga, 2016), part-of-speech tagging (Barrett et al., 2016, 2018b) and sentiment analysis (Mishra, Dey, and Bhattacharyya, 2017).

The most common representation of eye-gaze features to apply in NLP is a word-level feature. These features are concatenated along with the word and other linguistic features to predict a label in different supervised learning settings such as in Part-of-Speech tagging (Barrett et al., 2016), named entity recognition (Hollenstein and Zhang, 2019), sentiment analysis (Barrett et al., 2018a) and essay scoring (Mathias et al., 2018). However, using these token-level features implies that eye-tracking features must be available at training and test data to evaluate the model performance. The scenario might be impossible for most NLP datasets unless eye-tracking re-annotation is performed. To alleviate this problem, Barrett and Søgaard (2015) introduced type-level gaze features that aggregate the lexicon of word types by average overall occurrences of each type from the training data to allow the eye-tracking features to be used where eye-tracking features were not available in different evaluation dataset. These features were used to predict Part-of-Speech tagging and performed better than token-level features for discriminating grammatical functions.

Integrating eye-gaze features as input representation requires parallel eye-tracking annotation, which is costly to collect. Meanwhile, type-level features need parallel eye-tracking records at the test time to perform predictions. In practice, type-level features are still prone to missing values due to out-of-vocabulary problems in a disparate domain, such as between news scientific reports and fiction texts in domain adaptation scenarios. One possible solution is to simulate eye movement to generate eye-gaze features.

Nilsson and Nivre (2009) proposed to predict eye movement as a transition-based parser, assuming the words are read in sequences from a dataset. The task of a transition-based parser is to predict either fixations sequence or sets of discrete labels. For example, a sentence T= (Mary, had, a, little, lamb) where a number from 1 to 5 represents the word's index. The sequence of fixations is Mary–little–Mary–lamb is represented by F= (1,4,1,5), while the corresponding set is S(F) = 1,4,5 (Nilsson and Nivre, 2009). They use contextual (previous and current) token length, frequency, and distance to outside context as input features to transition-based models to predict fixations sequences. The features were inspired by early computational eye-modelling

research (Reichle, Rayner, and Pollatsek, 2003), where eye-movement patterns such as fixation were found to be correlated to the frequency and length of the tokens in normal reading situations.

Hahn and Keller (2016, 2018) use LSTMs to predict whether a word will be fixated or skipped. They tuned probability output from the LSTM language model from validation sets. It is achieved by thresholding the next word probability values. Other past studies (Barrett et al., 2018b; González-Garduño and Søgaard, 2017) use gaze estimation as an auxiliary task in multi-task learning settings. Their multi-task objectives are to predict text readability and unsupervised induction of Part-of-Speech tagging. They use gaze features as real values and regression loss as the auxiliary task. This allows the shared LSTM layer to improve the main task. It achieves better performance compared to other baselines. Another interesting representation is minimizing regression loss from the attention (hidden) layer output instead of the output layer to normalized gaze features to predict multiple tasks such as sentence-level classification (Barrett et al., 2018a).

These past studies demonstrate how eye gaze could be used with NLP models either for improving the models or as a task itself. In this study, I focused on evaluating the NLP models to the eye-gaze data related to the corresponding task of the model and the activity of human.

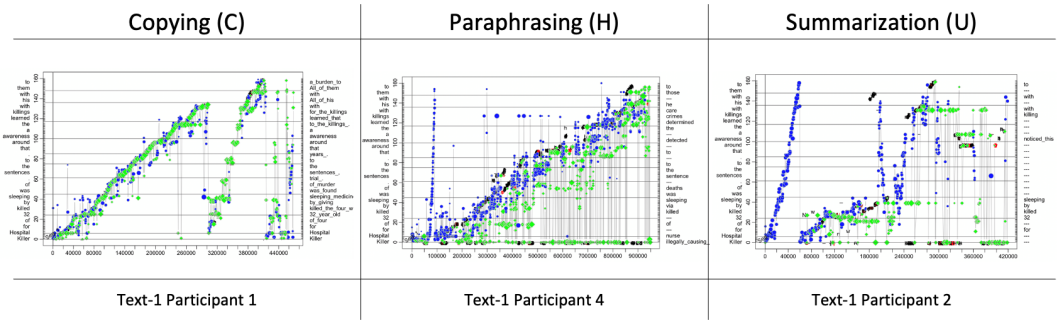


FIGURE 2.1: Eye-tracking and key-logger on three different writing tasks (Sahoo and Carl, 2019) .

2.1.3 Eye-gaze in Writing activity

A bulk of eye-gaze datasets in writing activity have been collected in translation process studies (Carl, 2012a). The collection is well supported by the data collection tool Translog(-II) (Carl, 2012b; Jakobsen, 1999; Schou, Dragsted, and Carl, 2009), which records user's eye-gaze points and keystroke logs. Although Translog was initially designed for the translation study, it can also be used for other writing activities (Sahoo and Carl, 2019). They collected three writing tasks: copying, paraphrasing, and summarising text. Figure 2.1 shows the fixation progress graphs of a text over three different tasks. Rodeghero and McMillan (2015) analysed eye movement patterns for program comprehension through writing in-line code summaries (Rodeghero and McMillan, 2015).

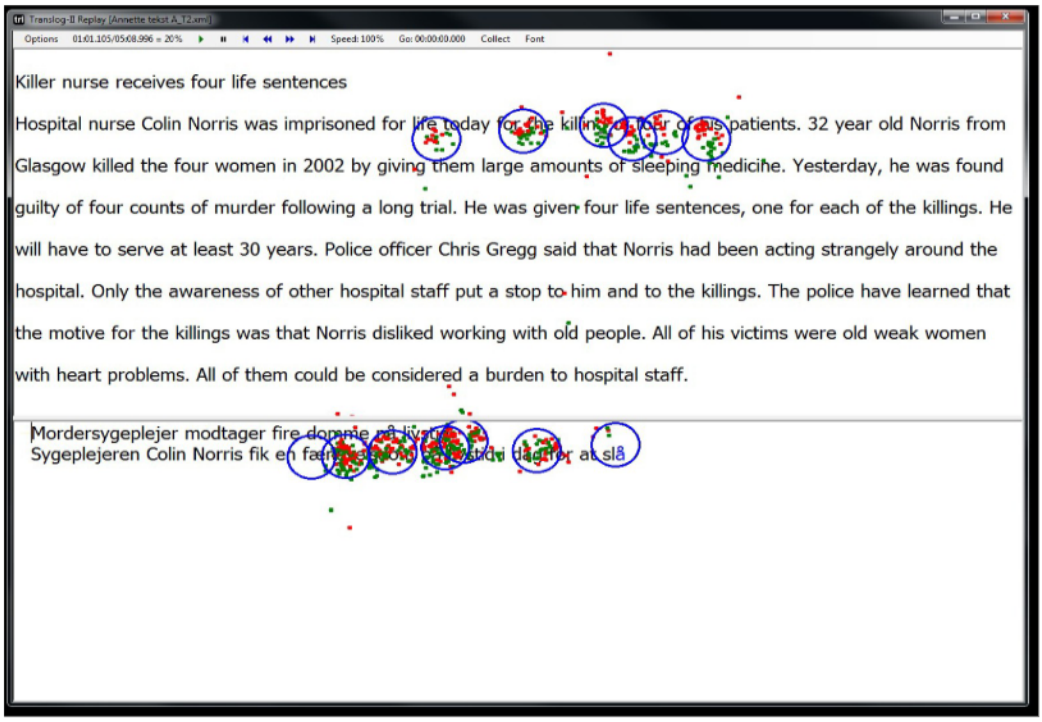


FIGURE 2.2: Translog-II interface with fixation circle visualization from recorded data using replay function (Carl, 2012b).

In this study, I use the Translog-II (Carl, 2012b) that records the user's eye movement from the eye-tracker device and keystrokes logs during a writing activity. It has been used to collect data in translation and monolingual studies mentioned earlier. Translog-II also transforms collected eye-gaze points into a sequence of fixations on words in the text on the screen. Figure 2.2 shows a screenshot of the software replaying the raw eye-tracking log. The eye-tracking log output format is an XML file.

2.2 Interpretability of Deep Learning models

The complexity of deep neural networks makes interpreting their behaviour inherently difficult. Most of the interpretation methods for deep neural networks resort to post hoc explanation (Ancona et al., 2018; Ribeiro, Singh, and Guestrin, 2016; Shrikumar, Greenside, and Kundaje, 2017; Simonyan, Vedaldi, and Zisserman, 2013; Sundararajan, Taly, and Yan, 2017) that extract the rationale of model outputs after the model has been trained. They are appealing due to their model-agnostic nature. For example, gradient as a post hoc interpretation measures the model sensitivity to input perturbation (Simonyan, Vedaldi, and Zisserman, 2013).

The computer vision (CV) field has actively studied the interpretation of neural networks. Zeiler and Fergus (2013) projects the intermediate convolutional layers onto an activation map using deconvolution to visualise the model activation. The projection from different layers shows the hierarchical nature of each CNN layer. For example, the shallow layers respond to corners and other types of edge/colour conjunctions. The middle layer captures similar textures, and the deeper layers capture class-specific information and characteristics of entire objects. A recent study found that the vision Transformer (ViT) is significantly less biased towards local textures compared to CNNs (Naseer et al., 2021). Simonyan, Vedaldi, and Zisserman (2013) proposed the saliency map that visualises the contribution of each pixel to the image-class score given an image and a class by computing gradient through back-propagation. They showed that the saliency map could be used for weakly supervised object localisation without requiring object bounding boxes or segmentation masks. The saliency map method then inspired other interpretation methods for CNN models, such as Local Interpretable Model-Agnostic Explanation (LIME), which removes subpixels (image region) from input image and learns the output probability differences (Ribeiro, Singh, and Guestrin, 2016), decomposing layer activation with local normalisation (Shrikumar, Greenside, and Kundaje, 2017) and integrated gradients (Sundararajan, Taly, and Yan, 2017). Motivated by the recent adoption of self-attention in ViT, Chefer, Gur, and Wolf (2021) proposed to combine the attention with relevance propagation (Binder et al., 2016) and evaluate the improved saliency with segmentation and perturbation metrics.

In the NLP field, model interpretation is not as straightforward as in the CV field. While intermediate features of CV (e.g. corners, edges and textures) are visible to humans, There are no such visible intermediate features except at the morphological level in the human language cognitive process. Instead, NLP interpretation studies (Jawahar, Sagot, and Seddah, 2019; Manning, Clark, and Hewitt, 2020) assume a layered structure of language leveraging linguistics concepts of morphology, syntax, semantics and pragmatics, which are theoretical and latent concepts. Moreover, the interaction between the linguistic layers is more complex than images. Straightforward compositionality, i.e., an element of a layer is made by combining elements of the previous layer and is less likely to hold in NLP than in CV.

Inspired by interpretation studies in the CV field, neural network interpretation has also been actively studied in NLP. Feng et al. (2018) measured neural models' sensitivity to input changes in reading comprehension tasks, such as textual inference and multi-modal QA. Their study revealed that the models tend to be overconfident and put higher saliency scores on the less meaningful part of the input sentences. For example, the models focused on punctuation and articles instead of relevant content words in the input sentence. They subsequently proposed a regularisation method to calibrate model *uncertainty* (the change of model prediction) by ablating input words.

In recent years, the investigation of the "attention" layer (Bahdanau, Cho, and Bengio, 2014) as an interpretation method has gained attention in the NLP community (Jain and Wallace, 2019; Serrano and Smith, 2019; Vashishth

et al., 2019; Vig, 2019; Wiegreffe and Pinter, 2019). An attention layer is a form of a matrix that represents the weight of interaction between different parts of inputs (Bahdanau, Cho, and Bengio, 2014; Chen et al., 2017; Luong, Pham, and Manning, 2015). It was inspired by human behaviour to select salient elements from stacks of information (Hassabis et al., 2017). There have been discussions of whether the attention layer could serve as an interpretation of a model prediction (Jain and Wallace, 2019; Serrano and Smith, 2019), or at least could be helpful to make sense of the model behaviour (Wiegreffe and Pinter, 2019). I also hypothesise that attention might provide an interpretation for NLP tasks (Wiegreffe and Pinter, 2019).

Study	Interpretation method	Architecture(s)	Task(s)
Hollenstein et al. (2019a)	Probing	Word embeddings, LSTM, Transformer	Intrinsic (regression)
Sood et al. (2020b)	Attention	LSTM, CNN, Transformer	Question-answering (QA)
Hollenstein and Beinborn (2021a)	Gradient, Attention	Transformer	normal reading (language model)
Eberle et al. (2022)	Attention flow	Transformer	sentiment analysis, relation classification
Brandl and Hollenstein (2022)	Attention	Transformer	bilingual reading
Pouw, Hollenstein, and Beinborn (2023)	Probing	Transformer	crosslingual
Maharaj et al. (2023)	Attention	CNN, Transformer	hallucination detection
Kuribayashi, Oseki, and Baldwin (2024)	Suprisal	LLM (Transformer)	normal reading
This work	Gradient, Attention, Input-perturbation	LSTM, CNN, Transformer	sentiment analysis, relation classification, QA, Summarization

TABLE 2.1: Studies on the interpretation methods for neural NLP models using eye-movement data

Several studies have investigated whether there is a similarity between the input word saliency provided by interpretation methods and that from human eye-gaze features. Here, I assume a possibility of correlation between the behaviours of deep learning models and human reading (Hale et al., 2018; Just and Carpenter, 1976). My thesis pursues this line of research. Whether eye-gaze features can explain the model performance in downstream tasks is also explored. Table 2.1 shows the comparison between the present work and existing studies using eye-gaze features. The goal of this thesis is similar to Hollenstein et al. (2019a) and Sood et al. (2020b); I investigate the relation between model and human behaviour. However, this thesis is more extensive because various interpretation methods are used on various model architectures across various tasks, from reading to writing activities. In this work, I assume that there are differences in human eye-movement behaviour across downstream tasks (Zelinsky et al., 2006), being triggered by the task goal. Hence, it is necessary to experiment with as many tasks as possible.

2.2.1 Interpreting models for Reading Task

An early study by Hollenstein et al. (2019a) compiled a collection of different modalities of cognitive data such as eye-tracking, EEG, and fMRI to evaluate word embedding semantic information. Sood et al. (2020b) focused on the attention layer in deep learning models and eye movements in question-answering (QA) tasks. Hollenstein and Beinborn (2021b) compared a language model and human behaviour by calculating a correlation between word importance from the model and that from human eye movements. Recent interests are also in analyzing multilingual model representation to language learner behaviour (Brandl and Hollenstein, 2022; Pouw, Hollenstein, and Beinborn, 2023). A recent study (Maharaj et al., 2023) applied the human attention bias to detect hallucination in the texts. Kuribayashi, Oseki, and Baldwin (2024) explores recent advances in LLM prompt methods to estimate human surprisal. However, the study still found that LLM prompt models underpredict human surprise compared to the baseline language model. This result highlights the relevancy and importance of building efficient NLP models from human cognition. In this thesis, I comprehensively investigated the models for various NLP tasks, such as sentiment analysis, relation classification and question answering (QA). All these studies target NLP tasks involving task-specific reading with eye-movement data.

2.2.2 Interpreting models for Writing Task

The autoregressive generation has been a popular technique for neural network-based language generation, where a word is generated according to the probability distribution over the vocabulary at each time step (Alvarez-Melis and Jaakkola, 2017; Vafa et al., 2021). The probability distribution is determined based on the network state at the previous time step. Unlike most reading tasks, which result in a single output, generation tasks involve a sequence of outputs (words). Therefore, there is a saliency distribution of input words at each time step, resulting in a saliency distribution matrix instead of a saliency vector.

The interpretation study for sequential generation has been active in the translation task (Alvarez-Melis and Jaakkola, 2017; Vafa et al., 2021; Voita, Sennrich, and Titov, 2021). The apparent application of the saliency score matrix is to analyse the alignment between source and target tokens in Machine Translation (Ding, Xu, and Koehn, 2019; He et al., 2019). This naturally can be used to calibrate attention models to improve performance (Lu et al., 2022). Recently, hallucination has been analysed with the model interpretation method (Tang, Fomicheva, and Specia, 2023; Xu et al., 2023).

Xu and Durrett (2021) investigated the role of the encoder and decoder components in the BART (Lewis et al., 2020) model in the summarisation task. Other studies in summarisation use text alignment for corpus creation (Tardy et al., 2020) and detect model hallucinations in summaries using mutual information (Poel, Cotterell, and Meister, 2022).

In this thesis, I propose analyzing the generation process of summarisation models with eye-gaze information. This is the first comparative study

of sequential models with human eye movement behaviour. I approach the saliency output from sequential models by transforming the vectors compatible with input-level saliency. I also propose a new representation of the eye-gaze data to analyse the feature in the model's decoding process.

Chapter 3

Interpreting Models in Reading Tasks

In this chapter, I describe my investigation into interpreting deep learning models in three NLP tasks related to a task-specific reading on human behaviour using eye-gaze data. The objective of this study is to determine the intricate interaction between the type of models in NLP, interpretation methods, reading tasks, eye-gaze features, and evaluation metrics in analyzing deep learning models related to the human gaze.

To answer RQ1 in reading tasks, I define a metric, "Saliency distance" (SD), comparing the saliency distribution over input words obtained from an interpretation method with an eye-gaze feature. This question holds significant practical implications. If machines exhibit different behaviour from humans and fail to achieve human-level performance, it indicates that machines can learn from human behaviour, inspiring the development of more human-like machine learning models.

To answer RQ2, I propose a novel evaluation method between model interpretation and eye gaze. The technique represents the cumulative model performance against the SD scores; I called it the "SD-performance curve" (SDPC). If the answer to RQ2 is affirmative, we cannot only identify the difference in behaviour between humans and machines but also pave the way for improving machine performance by training them to emulate human behaviour.

3.1 Interpretation methods

Interpretation for deep neural networks is defined as assigning a saliency score to each input element with respect to a neural model prediction (Ancona et al., 2018). The saliency scores suggest the importance of input elements in the model's decision making. Formally, given an input text X consisting of N words, the text is passed through a neural model f and gets its prediction class c. The saliency score $\phi(x_i, c)$ is assigned for each input word x_i. In NLP tasks, a word x_i is often represented as a real vector $\vec{v_i} \in \mathbb{R}$ obtained from an embedding layer. The saliency score for a word $\phi(x_i, c)$ is, therefore, in the form of a vector, expressing the saliency of an element in the word embedding. Then, this vector is converted into a scalar saliency score

for each word; different conversion methods are used for different interpretation methods. Therefore, the saliency score for an input text is in the form of a vector. This vector is compared with eye-gaze features (explained in Section 2.1). The following subsections explain various post hoc interpretation methods to calculate the input word saliency score.

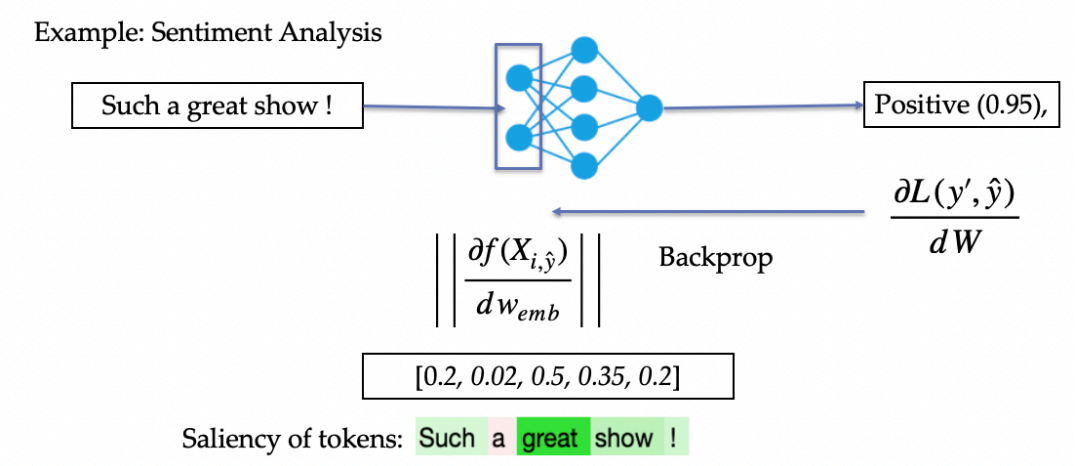


FIGURE 3.1: Gradient based saliency.

3.1.1 Simple gradient (Grad)

The gradient-based interpretation method calculates the saliency scores of input elements via back-propagation. This method was initially proposed for CNN, also commonly known as heat-map. Formally, the saliency score $\phi_{\text{Grad}}(x_i, c)$ for an input element x_i is obtained from the gradient $\frac{\partial f(X)}{\partial x_i}$. In NLP tasks, the saliency score for each input word x_i is in the form of a gradient vector for its corresponding embedding $\vec{v_i}$. The norm of the gradient vector is then used to represent the saliency score of the word. Figure 3.1 illustrate how to obtain gradient saliency with backward propagation.

3.1.2 Integrated gradient (IG)

Shrikumar, Greenside, and Kundaje (2017) revealed that the simple gradient method underestimates the importance of the input. This is caused by the saturation of non-linear activation functions in deep neural networks. Sundararajan, Taly, and Yan (2017) proposed to alleviate this problem by integrating the gradients using a linear interpolation over $\alpha \in [0, 1]$. The integrated gradient is computed between a *baseline* input $\hat{X}$ and the original input X (Ancona et al., 2018).[1] It is computed as in Equation (3.1), where b_i

[1]Refer to Ancona et al. (2018) for the *baseline* input.

is an element of B.

$$\phi_{\text{IG}}(x_i, c) = \int_{\alpha=0}^{1} \frac{\partial f(B)}{\partial x_i} \partial\alpha \qquad (3.1)$$
$$B = X + \alpha(X - \hat{X}), x_i, \in X \text{ and } b_i \in B$$

As in the simple gradient, the norm of the $\phi_{\text{IG}}(x_i, c)$ is computed on the word vector representation $\vec{v_i}$ for each input word x_i to represent its saliency score.

3.1.3 Input-perturbation (IP)

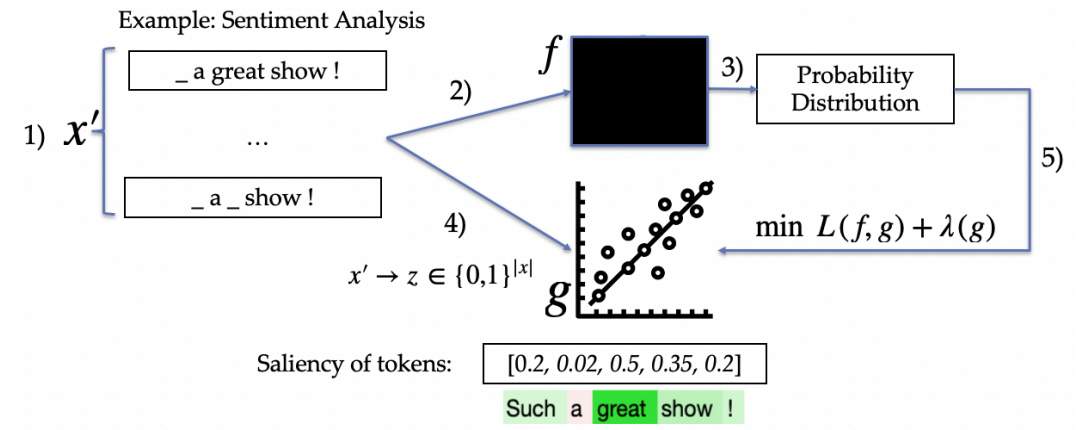


FIGURE 3.2: Input perturbation based saliency.

While the whole input X is passed into the neural model in the gradient-based methods, the input-perturbation method utilises a part of the input X. Specifically, Local Interpretable Model-agnostic Explanation (LIME) (Ribeiro, Singh, and Guestrin, 2016) is employed in this research. LIME operates by sampling the neighbourhood of interpretable components locally and then uses a linear surrogate function to represent a given classifier. For example, to assign the saliency score of an input element x_i, the input $\hat{X}$ is passed to an approximation model $g(\hat{X})$ while removing x_i from X. Formally, LIME uses a linear model g to approximate the behaviour of the original neural model $f(X)$. The saliency score vector $\phi_{\text{LIME}}(X, c)$ is then calculated as in Equation (3.2). The linear model minimise the loss between the surrogate function g and the original function f while keeping the complexity of the surrogate function low using a regularisation term λ.

$$\phi_{\text{LIME}}(X, c) = \min_{g} \mathcal{L}\left(f(X), g(\hat{X})\right) + \lambda(g) \qquad (3.2)$$

Figure 3.2 shows the step-by-step processes of obtaining LIME method saliency.

3.1.4 Attention (Attn)

Attention is one of the widely used modules in the encoder-decoder model for neural machine translation (NMT) (Bahdanau, Cho, and Bengio, 2014). It

represents the alignment between words in source and target sentences. It is also beneficial in the natural language inference (NLI) (Parikh et al., 2016) and QA tasks (Seo et al., 2017; Xiong, Zhong, and Socher, 2016), which take a pair of texts as an input (premise–hypothesis pairs in NLI, context–question pairs in QA). The attention layer outputs a matrix representing the pairwise saliency scores between words across the input texts.

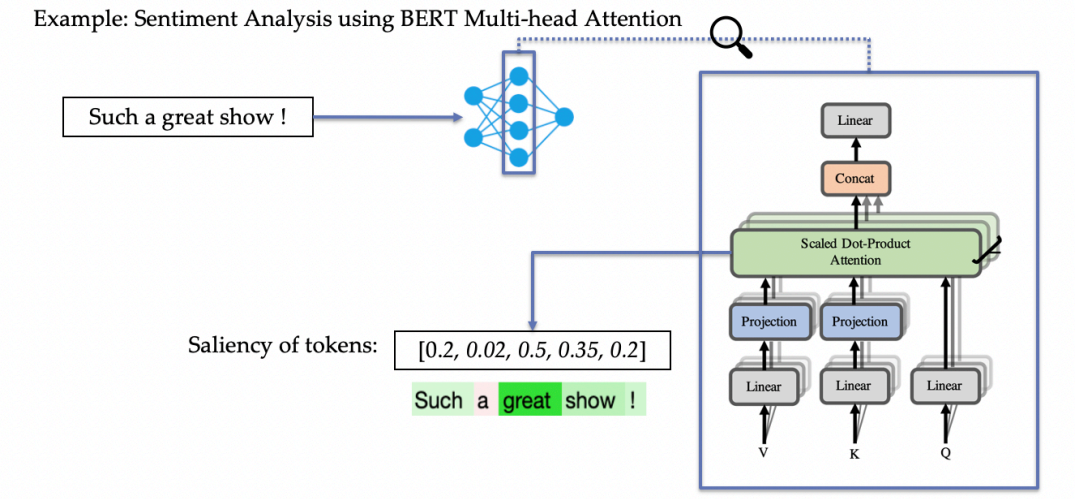


FIGURE 3.3: Attention based saliency which extract attention weights, usually from last attention layer of a trained model.

In general, attention is computed as a pairwise conditional distribution using a similarity function $\Phi_{\text{att}}(h, q)$, where h is a context vector (e.g., a source sentence in NMT, a hypothesis in NLI and a context in QA) and q is a query vector (e.g., the previously generated word for the output target sentence in NMT, the premise in NLI, the question in QA). The outputs of the similarity function are normalised scalar scores between the context and query, typically using the Softmax function. There are several methods to compute the attention score. In this work, I use a dot product $\Phi_{\text{dot\_att}}(h, q) = \text{softmax}(hq)$ (Luong, Pham, and Manning, 2015) for LSTM and CNN and scaled dot-product $\Phi_{\text{sca\_dot\_att}}(h, q) = \text{softmax}\left(\frac{hq}{\sqrt{m}}\right)$ (Vaswani et al., 2017) for Transformer (Figure 3.3), where m is the dimension of h and q.

3.2 Eye-gaze Dataset in Task-Specific Reading

I use two datasets that contain eye-movement data when humans solve NLP tasks. These two datasets are publicly available with human annotation and their gold standard label. The datasets comes with gold standard label because the texts stimuli are come from published corpora.

3.2.1 Task-Specific Reading: ZuCo

ZuCo (Hollenstein et al., 2018, 2019b) is a publicly available eye-movement dataset containing human eye fixations in three NLP tasks: sentiment analysis, relation classification and multiple-choice question answering. Eye movement was recorded from 12 English native speakers who participated in solving the above three tasks.

In the sentiment analysis task, the participants are asked to evaluate the sentiment of movie reviews on a 1 (negative)–5 (positive) scale. The sentiment analysis data come from the Stanford Sentiment Treebank (SST) (Socher et al., 2013). In the relation classification task, the participants judge whether one of the predefined 11 relations holds between entities in the input sentence (Culotta, McCallum, and Betz, 2006). Participants do not need to extract the entities that the relation holds. In the question-answering task, the participants are asked to select the correct choice given a question after reading a context sentence from Wikipedia texts. The correct choices are verbatim excerpts from the context sentences.

The ZuCo data includes human eye-gaze information when the participants solve the above tasks. Various tasks enable us to analyse the difference in reading strategies in different tasks (Poole and Ball, 2006). The statistics of the ZuCo dataset are shown in Table 3.1; Column #sent denotes the number of sentences in the task, which is the same as the number of question instances and #token is the total number of tokens in the sentences. Column #annotation means the number of test instances annotated with the participant answers. I use the test instances with annotation in the experiments.

Task	#sent	#token	tokens/sent	#annotation
Sentiment analysis	400	7,079	17.70	45
Relation classification	407	8,164	20.06	385
Question answering	300	6,386	21.29	55

TABLE 3.1: Statistics of the ZuCo dataset for the sentiment analysis, relation classification and question answering tasks.

3.2.2 Multiple Choice Question Answering: MQA-RC

MQA-RC (Sood et al., 2020b) contains human eye-gaze information when answering a part of questions in the MovieQA dataset (Tapaswi et al., 2016). A question item of MovieQA consists of a synopsis (200–250 words), a question and five answer choices. After reading the synopsis and question, the participants were asked to select a correct choice. The MQA-RC dataset includes 32 question items; each question item was answered by five, six or 16 participants according to the experimental designs. I use all data from the MQA-RC dataset.

3.3 Eye-gaze Dataset Analysis

In this section, I analysed eye-gaze features similarity between participants. This analysis aims to establish the upper-bound agreement on the ZuCo and Movie-QA datasets. Another objectives is to analyse which features should be compared with models to answer the research questions.

3.3.1 ZuCo Dataset

First, I analyse the inter-agreement of each eye-gaze feature, such as Fixation Count (FC), First Fixation Duration (FFD), Gaze Duration (GD), and Go-Past-time (GPT) features in the ZuCo dataset. Kendall-tau is used to measure the relative ranking similarity between participants. I calculate the inter-agreement similarity in two methods.

1. Pairwise: I calculate a pairwise similarity between the target participant and each of the rest participants. I report the average of the similarity of all pairs.

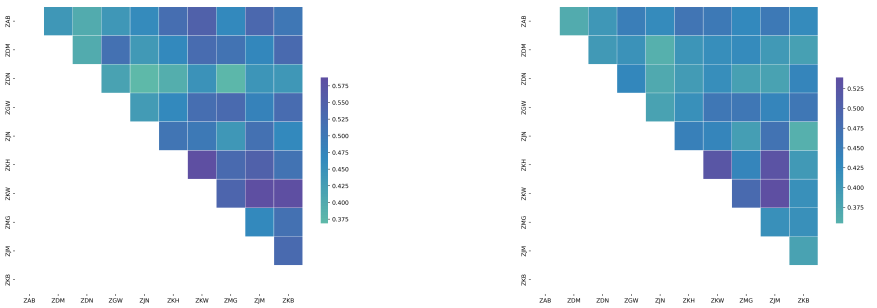
2. Leave-One-Out: I calculate the average of features of participants except for the target participant for each text. Then, I calculate the similarity between the target participants and the average. This approach expects the average eye-gaze features are more reliable than individual features.

	FC	FFD	GD	GPT
Sentiment Analysis (SA)				
Pairwise	0.453	0.304	0.351	0.385
Leave-One-Out	0.619	0.498	0.539	0.541
Relation Classification (RC)				
Pairwise	0.389	0.290	0.314	0.314
Leave-One-Out	0.576	0.491	0.510	0.491
QA-Span (Normal Reading)				
Pairwise	0.448	0.303	0.355	0.395
Leave-One-Out	0.620	0.500	0.544	0.554

TABLE 3.2: Gaze features correlation between Subjects in ZuCo datasets for each tasks (Sentiment Analysis, Relation Classification and QA-Span). Results is mean of multiple Kendall-tau between pairs of input for each subjects.

I report the pairwise and leave-one-out similarity for the three tasks and four eye-gaze features in Table 3.2. From the table, It can observed that FC features provide higher inter-agreement than other features. Specifically, the leave-one-out FC features in other subjects* correlate moderately (>0.6). In contrast, the FFD feature provides the lowest inter-agreement compared to other features. Still, FFD might be beneficial because it represents a one-pass left-to-right fixation duration and does not include complex features such

as regression. Figures 3.4 show two heatmaps (Subfigure 3.4a and 3.4b) for sentiment analysis and relation classification tasks. The heatmap visualises a breakdown of details of the inter-agreement similarity between participants.



(A) Sentiment Analysis eye-movement inter-subject FixCount correlation

(B) Relation Classification eye-movement inter-subject FixCount correlation

FIGURE 3.4: Inter-agreement heatmap between subjects in ZuCo sentiment analysis and relation classification tasks

3.3.2 Movie Plot Question Answering

Sood et al. (2020b) collect The MQA-RC with two different scenarios:

1. The scenario 1 experiment was conducted in three different conditions for the first scenario.

   (a) The first experiment (exp. 1) was a default scenario where the participants could access the plot, questions, and choices.

   (b) The second experiment (exp. 2) was a free-form QA where multiple choices were not used, but the participants had to write or select the answer from the passage.

   (c) The last experiment (exp. 3) was called a memory condition where the participants had to read the plot separately from questions and choices.

   In scenario 1, the participants were divided into three different schemes: A (exp 1, 2, 3), B (exp 2, 3, 1), and C (exp 3, 1, 2), where participants only performed each experiment once and in different orders.

2. Scenario 2 is a follow-up study of the first experiment (Sood et al., 2020b). The goal is to compare selected documents where Hierarchical CNNs and LSTMs (Blohm et al., 2018) predict correct answers from MovieQA validation set.

I analyse the inter-agreement between participants' fixation count. The analysis aims to measure how similar the focus between participant. Figure 3.5 shows the pairwise heat-map between pair of participants for each

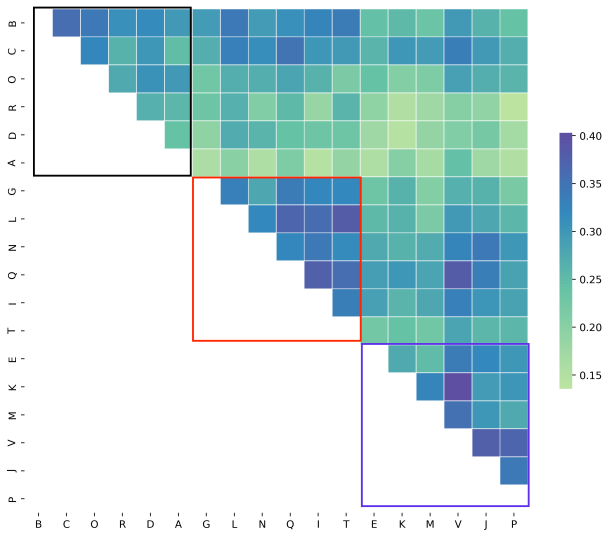


FIGURE 3.5: MovieQA Eye-movement dataset inter-gaze agreement

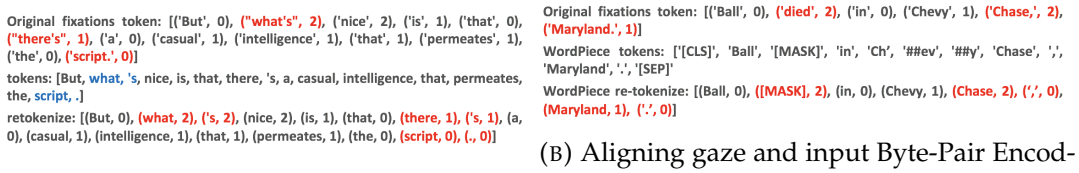
	schema A	schema B	schema C
schema A		-0.369	2.083*
schema B			-2.833†
schema C			

TABLE 3.3: Pairwise differences between schemas. * means p-value < 0.05, †means p-value < 0.01.

cell. From the figure, some patterns between the schemes can be observed. Furthermore, I measure the significant difference between participants in different schema. The aim is to test whether the task order causes a significant difference in eye-gaze pattern between groups on different task orders of the three schemes above on Table 3.3. From Table 3.3, it is observed that schema C (exp 3, 1, 2) participants differ significantly from schema A (exp 1, 2, 3) and B (exp 2, 3, 1). I hypothesise that is mainly due to the condition given the passage first and having to recall a passage from memory to answer multiple-choice QA.

3.4 Eye-gaze features

This section explains the eye-gaze features to compare with model saliency. I focus on two key eye-gaze features, namely, *fixation count* (FC) and *first fixation duration* (FFD), which have significant implications in the fields of cognitive science and human-computer interaction. FC is a cumulative count of



(A) Aligning gaze and input token

(B) Aligning gaze and input Byte-Pair Encoding

FIGURE 3.6: post-processing to realign tokenization done in gaze features (white-space tokenizer) to subfigure(A) standard English tokenizer (punctuation) and subfigure(B) BPE

fixations over an AOI during the entire task. In contrast, FFD is the first fixation duration over an AOI. While FC is concerned with the eye gaze on the AOI during the entire reading activity, FFD is concerned only with the first look at the AOI. The past studies (Hollenstein and Beinborn, 2021a; Sood et al., 2020b) typically adopted FC or total fixation duration as the eye-gaze feature. To my knowledge, no previous study has used FFD to compare with interpretation methods. While FC considers all of the fixations, FFD extracts features from the first reading run of the text. Since neural models read through an entire input text simultaneously, whereas humans can read the text in several passes, I assume that FFD might provide new observations in analyzing differences between human and machine behaviours.

3.5 Tasks and experimental settings

This section discusses the definition of each task, followed by the models, their implementation details and the evaluation metrics used in the experiment. I consider three architectures, LSTM, CNN and Transformer, and four interpretation methods, Grad, IG, IP and Attn. For all tasks, the reading experiments use CNN, LSTM and Transformer-based models. I implemented the attention mechanism in the LSTM and CNN-based models only when evaluating the Attn interpretation method. I used all four interpretation methods for the sentiment analysis and relation classification tasks. However, due to lengthy inputs and outputs, I did not apply the IP interpretation method for the question-answering tasks.

In this study, two eye-gaze features are considered: *fixation count* (FC) and *first fixation duration* (FFD).

3.5.1 Sentiment analysis (SA)

The sentiment analysis is a task to predict a sentiment score of an input sentence; the score ranges from one (very negative) to five (very positive).

Given an input sentence as a sequence of word embeddings $\overrightarrow{v_i}, ..., \overrightarrow{v_N}$, The input sentence is encoded into a single vector representation $\overrightarrow{s}$ by using three types of architectures: LSTM, CNN, and Transformer. The LSTM's hidden layer output from the last time-step is retrieved as the sentence representation. In the LSTM encoder, The hidden dimensions of 512 was used.

In CNN, the sequence of embeddings is pooled using the maximum pooling to create the sentence representation. The CNN encoder consists of four channels (1,3,5,7) with 192 dimension outputs each. Both CNN and LSTM encoder accept 300 dimension inputs from Glove pre-trained word embedding (Pennington, Socher, and Manning, 2014). In this experiment, I employ the `bert-base-cased` pre-trained model (Devlin et al., 2019) as the Transformer architecture. The [CLS] vector is used for the sentence representation. The sentence representation of each architecture is passed to a Softmax layer to get the final sentiment label prediction. The experiment was run five times with different random seeds.

The original SST (Socher et al., 2013) train split (8.5K sentences) is used for training the LSTM and CNN models from scratch and for fine-tuning the BERT model. In training the models, the Adam optimiser (Kingma and Ba, 2017) with a learning rate of 0.001 and a batch size of 32 was used. As testing data (i.e. comparing models' behaviour with human eye-gaze features), this experiment use 45 annotated sentences from the ZuCo dataset (Table 3.1).

3.5.2 Relation Classification (RC)

Given an input sentence containing two entities, an RC model predicts which relation out of 11 holds between them. Two entities in the input sentence are not marked, which would be implicitly identified by the models for identifying their relation. However, the models do not have to output the entities.

This experiment use the same three architectures as the SA task: LSTM, CNN and Transformer (BERT). Most of the training settings remain the same as the sentiment analysis models. I train the models using the Adam optimiser with a learning rate of 0.001 and a batch size of 64. I set a dropout value to 0.5 (between embedding–encoder layer, and encoder–classifier layer) to prevent over-fitting due to the small data size.

I use the training split of Culotta, McCallum, and Betz (2006)'s dataset (1K sentences), which does not overlap with the ZuCo dataset. In this experiment, I train five models for each architecture using five-fold cross-validation. In every fold, 55 sentences from the ZuCo-RC dataset (Table 3.1) are used for testing, i.e. comparing models' behaviour with the eye-gaze features. The experimental results was obtained by averaging scores from the five models.

3.5.3 Question Answering (QA)

There are two kinds of QA tasks prepared: span-based and multiple-choice QA tasks. Due to the high computational cost, I do not consider the input-perturbation method for the QA task. Unlike the other tasks, the output of the QA-Span is a sequence of words instead of a single class, such as a sentiment scale (SA), a relation between entities (RC) or a correct choice (QA-MC). Therefore, many ablation patterns should be considered for the IP method. Also, as shown in Table 3.1, the input texts of the QA-MC task are longer than the other tasks, leading to many ablation patterns.

Span-based QA (QA-Span) ZuCo's question-answering task is a multiple-choice QA on Wikipedia texts. However, a large dataset for the multiple-choice QA on Wikipedia is unavailable. As described in section 3.2, ZuCo's answer choices are verbatim excerpts from the context sentences. Using this characteristic, the ZuCo's multiple-choice QA is transformed into a span-based QA task by identifying the answer span in the context sentences by referring to the answer choice. The input to the model for testing is a pair of a context sentence and a question; the output is the start and end tokens of the answer span in the context. The SQuAD dataset (85.6K QA-pairs; Rajpurkar et al. (2016)) is used, which was built for the span-based QA task on Wikipedia text, for training the models. For testing with the eye-gaze features, I only used 55 instances with annotation from the ZuCo dataset of the QA task (Table 3.1).

BiDAF (Seo et al., 2017) is used for the LSTM architecture, QANet (Yu et al., 2018) for the CNN architecture and the BERT-based QA model (Devlin et al., 2019) for the Transformer architecture.

Multiple-choice QA (QA-MC) I use the training split of MovieQA (Tapaswi et al., 2016) dataset (9.8K QA-pairs) as the training set in the multiple-choice QA task, and use 32 instances with eye-gaze features from MQA-RC (Sood et al., 2020b) as the testing set.

The LSTM and CNN-based models (Blohm et al., 2018) take a triplet of ⟨source text, question, multiple choices⟩ as input. The model answers the question by selecting the most probable option. The XLNet-large (Yang et al., 2019) model is used for Transformer using the HuggingFace implementation (Wolf et al., 2020), and fine-tune the model with the training set. The model takes five triplets of ⟨source text, question, one of the choices⟩ as input where the source text and question are shared across the five triplets. The experiment is ran ten times with different random seeds.

3.5.4 Post-processing

Most eye-tracking data tokenization methods use white space as a separator following the visual paradigm. To analyse the input attribution score of a model given a predicted input to fixation data, which used different tokenization methods like punctuation for Word-based embedding input or Byte-Pair Encoding (Sennrich, Haddow, and Birch, 2016) for BERT. I post-process fixation tokens using a punctuation-based tokenizer and duplicate the feature if the content is a word or zero if the element is just punctuation or a nonsensical symbol. For example, a word contraction ["who's"] is tokenized into ["who", "'", "s"], in all of our evaluation scenarios. The post-process aligning between the gaze token and model input token can be matched one-to-one by "traversing" the tokenization process and tracking the same length given the same tokenizer method, as illustrated in Figure 3.6a. In byte-pair-encoding-based tokens, I merge sub-words input representation into a token

by the merging scenario and summing[2] the corresponding index of attribution score (figures 3.6b).

3.6 Results for RQ1: Does the input word saliency from interpretation methods conform with human eye-gaze features?

3.6.1 Saliency distance (SD)

To answer RQ1, I propose to compare the saliency distribution over input words obtained from an interpretation method (P) with that from an eye-gaze feature (Q). The min-max normalisation is applied to the word saliency values from the interpretation methods before calculating the distribution over the input words. When calculating the eye-gaze features, fixations from different participants are not distinguished. Instead, the average feature values across all participants are used for each test instance to make a single distribution of the feature over the input words.

Following Sood et al. (2020b), I also measure their conformity using KL-divergence $\mathrm{KL}(P\|Q)$ (Kullback and Leibler, 1951). This technique is applied to compare the similarity of two probability distributions into a relative entropy. The relative entropy will also be easier to deal with when optimising a model in learning algorithms instead of ranking correlation. The KL divergence is computed between the average human and the average model along the input sequence of words. Before comparing into KL, both model saliency and human gaze are normalised. Each sequence of model saliency values and human eye-gaze feature values is summed. Then, saliency and eye-gaze feature values on tokens are normalised so that they sum up to one. We consider these values a distribution across tokens. This handle different peaks in the data into a smooth and relative distribution within the same range.

A smaller KL-divergence value indicates a higher similarity between distributions, meaning that a model focuses on the words which humans look at to achieve the task. The KL-divergence here is used to compare between models that have high similarity or low relative entropy to human. In the remainder of this thesis, The KL-divergence value is referred to as the saliency distance (SD) between the interpretation method and the eye-gaze feature. Table 3.4 shows average SD scores for the combinations of the tasks, interpretation methods, and architectures. Columns FC and FFD denote the eye-gaze feature used for calculating SD scores. Numbers in boldface denote the smallest values for the architecture in the same setting.

Sentiment Analysis (SA)

The Grad interpretation method with Transformer shows the smallest SD

[2] I test sum, averaging, and max merging of the attribution score. Empirically, simple sum performs best in all interpretation methods

Interpretation method	Task / Architecture	SA FC	SA FFD	RC FC	RC FFD	QA-Span FC	QA-Span FFD	QA-MC FC	QA-MC FFD
Grad	LSTM	1.018	0.963	2.497	2.343	1.091	1.058	2.003	1.969
	CNN	0.761	0.735	1.122	1.061	0.985	0.951	1.014	0.987
	Transformer	**0.667**	**0.656**	**0.779**	**0.752**	**0.800**	**0.778**	**0.489**	**0.410**
IG	LSTM	0.899	0.848	2.342	2.195	1.015	0.991	1.047	1.012
	CNN	0.793	0.776	1.105	1.054	1.066	1.028	0.996	0.953
	Transformer	**0.712**	**0.699**	**0.829**	**0.794**	**0.922**	**0.887**	**0.469**	**0.390**
IP	LSTM	**2.619**	**2.459**	2.968	2.855				
	CNN	3.317	3.137	3.073	2.921				
	Transformer	4.028	3.686	**2.825**	**2.728**				
Attn	LSTM	2.597	2.584	2.436	2.389	2.622	2.635	**0.403**	**0.328**
	CNN	2.507	2.649	**1.898**	**1.895**	3.114	3.018	0.444	0.373
	Transformer	**1.385**	**1.484**	3.002	2.901	**0.336**	**0.297**	0.466	0.387

TABLE 3.4: Saliency distance (SD) for the combinations of task, interpretation method, architecture and eye-gaze features. Bold denotes the smallest value of the architecture with the same setting.

scores among all combinations of the interpretation methods and the architectures. Transformer shows the smallest SD scores with the Grad, IG and Attn methods except for the IP method. The IP method always shows larger SD scores than other methods, particularly with Transformer. Lastly, the SD scores by FFD are generally smaller than those by FC, except for the Attn method with CNN and Transformer.

Relation Classification (RC)
A similar tendency is observed in the SA task for the gradient-based interpretation methods. The Grad and IG methods show smaller SD scores than the Attn and IP methods, with notable differences. The Grad, IG, and IP methods with Transformer show the smallest SD scores, while the Attn method shows the smallest SD score with CNN instead of Transformer. All SD scores by FFD are smaller than those by FC in this task.

Span-based QA (QA-Span)
Transformer consistently shows smaller SD scores than the other architectures across all interpretation methods, particularly the smallest for the Attn method. The SD scores by FFD are smaller than those by FC except for the Attn method with LSTM.

Multiple-choice QA (QA-MC)
The gradient-based interpretation methods (Grad and IG) consistently achieve the smallest SDs with Transformer, the same as for the other three tasks. On the other hand, the Attn method shows the smallest SD scores with LSTM

instead of Transformer, although the differences are small. All SD scores by FFD are smaller than those by FC for the QA-MC task.

3.6.2 Discussion for RQ1

Transformer tends to show the smallest SDs across tasks in terms of architecture. The exceptions are when using the IP method for the SA task, the Attn method for the RC and QA-MC tasks.

From the viewpoint of the interpretation method, the gradient-based interpretation methods (Grad and IG) show smaller SD scores with both eye-gaze features for the SA, RC and QA-Span tasks, while the Attn method shows smaller SD scores for the QA-MC task. The main difference between the SA, RC and QA-Span tasks and the QA-MC task is the text length (cf. Table 3.1); a sentence (around 20 words) is an input for the formers, while a lengthy text (more than 200 words) and a question are an input for the latter.

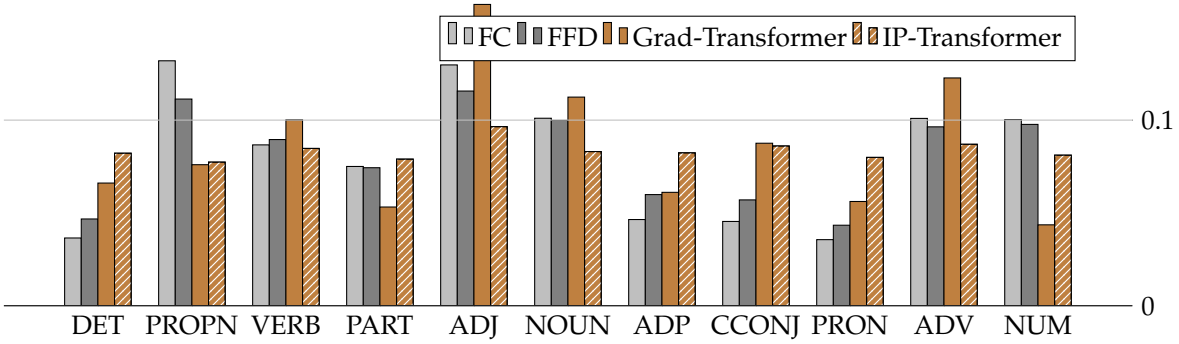


FIGURE 3.7: PoS saliency distribution of the eye-gaze features and the best (Grad-Transformer) and worst (IP-Transformer) combinations of the interpretaion method and architecture for sentiment analysis (SA)

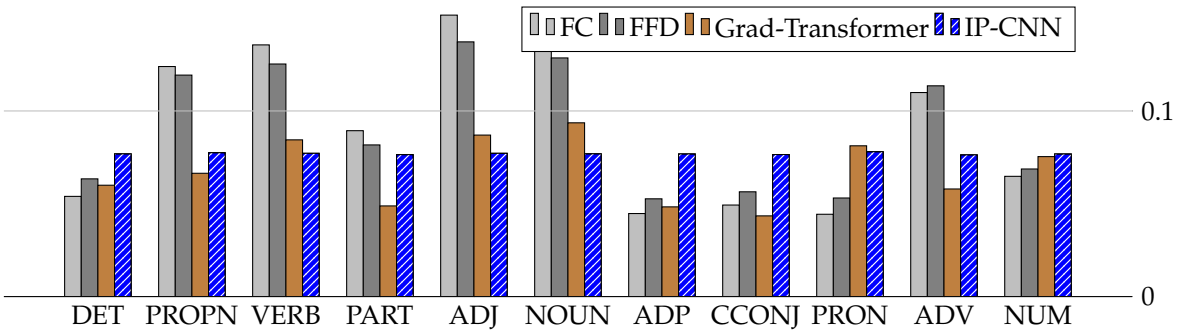


FIGURE 3.8: PoS saliency distribution of the eye-gaze features and the best (Grad-Transformer) and worst (IP-CNN) combinations of the interpretaion method and architecture for relation classification (RC)

For a more detailed analysis, I calculated the saliency distribution over Parts of Speech (PoS) for the SA and RC tasks. I applied a PoS tagger [3] to the input sentence, summed up the word saliencies according to their PoS each, and normalised the PoS saliencies to sum up to 1.0. Considering the task goals, I presume that sentiment-suggesting PoS like nouns, adjectives and adverbs attract more attention in the SA task, and PoS like nouns, verbs, and adjectives do in the RC task, where nouns correspond to entities, and verbs and adjectives suggest their relations.

Figure 3.7 shows the PoS saliency distribution of the eye-gaze features (FC and FFD) and the best and worst conformity combinations for the SA task. According to Table 3.4, the combination Grad-Transformer shows the best conformity, i.e. the smallest SD score, and the IP-Transformer shows the worst conformity, i.e. the largest SD score. As presumed, humans look at proper nouns (PROPN), adjectives (ADJ), nouns (NOUN), adverbs (ADV) and numbers (NUM). The best model focuses on ADJ more than humans but on PROPN and NUM less than humans. The best model also focuses on determiners (DET) and coordinating conjunctions (CCONJ), which are not helpful for sentiment analysis intuitively. In contrast, the focus of the worst model is relatively uniform compared to that of the best model.

Figure 3.8 shows the PoS saliency distribution for the RC task. In the RC task, humans look at PROPN, VERB, ADJ, NOUN and ADV. In contrast with the SA task, the models do not capture this tendency well. Notably, they do not focus on VERB, ADJ and ADV, which can indicate relations between entities. Again, the worst model focuses on every PoS uniformly.

For the QA tasks, the saliency of words inside and outside the answer spans in the context sentence is compared. Table 3.5 shows average word saliencies inside (IN) and outside (OUT) the answer span each. The Diff column shows the differences between the IN and OUT column values, i.e., the IN value minus the OUT value. The asterisk indicates statistical significance at a .05 significance level by the paired permutation test.

For the QA-Span task, both humans and models focus more on IN words than OUT words except for the Attn method. Although the Attn method shows small total SD scores for the QA-Span task in Table 3.4, there are differences through this IN-OUT word saliency analysis.

Hollenstein and Beinborn (2021a) reported that the gradient-based method shows more similar tendencies to humans than the Attn method when using the Transformer architecture for reading the ZuCo QA text. Their results contradict the QA-Span results in Table 3.4 but conform with this detailed analysis (Table 3.5). In addition to the task difference (text reading vs. span-based QA), other differences must also be noted in the experimental setting. Although the data is the same, my data is a subset of theirs; they used the fixation duration, while I used the fixation count and the first fixation duration.

Concerning the QA-MC task, the overall result (Table 3.4) agrees with the findings by Sood et al. (2020b). They reported that the hierarchical attention LSTM-based model is more similar to humans than the Transformer-based model (XLNet) for the QA-MC task when using the Attn method. However,

[3] Spacy (Honnibal and Johnson, 2015)

	Task	QA-Span			QA-MC		
		IN	OUT	Diff	IN	OUT	Diff
Eye-gaze	FC	.070	.059	.011*	.007	.005	.002*
	FFD	.071	.059	.012*	.006	.005	.002*
Interpretation method	Architecture						
Grad	LSTM	.118	.053	.065*	.005	.005	.000
	CNN	.123	.052	.071*	.005	.005	.000
	Transformer	.135	.048	.087*	.005	.006	−.001
IG	LSTM	.124	.052	.072*	.005	.005	.000
	CNN	.156	.046	.110*	.005	.005	.000
	Transformer	.129	.051	.078*	.005	.006	−.001
Attn	LSTM	.061	.061	.000	.005	.005	−.000
	CNN	.061	.061	−.000	.006	.005	.001*
	Transformer	.061	.060	.000	.004	.006	−.002

TABLE 3.5: Average saliency values over inside and outside of answer spans.
* denotes statiscial significance at a significance level p<.05 by the paired permutation test.

the detailed analysis in Table 3.5 does not support this claim. There is a significant difference in saliencies inside and outside the answer spans for the Attn-CNN combination but not for other models and interpretation methods combination. While they tested only the Attn method, I also tested the two gradient-based methods. The saliency difference between IN and OUT words cannot be observed using the gradient-based method either.

Lastly, Table 3.4 shows that SD scores by FFD are mostly smaller than SD scores by FC even though the overall tendencies are similar. This tendency might be explained by the fact that FC considers multiple looks at words. In contrast, FFD considers only the first look at words, which is similar to the model's behaviour, i.e. machines do not look back at the same word multiple times. Some might argue that deeper layers in deep learning neural networks might capture the re-reading mechanism. Investigating the deeper layers of the networks is left as future work. Based on this finding, it is decided to use FFD to analyse the relationship between the model saliency and the model performance to answer the second research question (**RQ2**) in the next section.

Interpretation method	Architecture	SA Success	SA Fail	RC Success	RC Fail	QA-Span Success	QA-Span Fail	QA-MC Success	QA-MC Fail
Grad	LSTM	1.018	0.906	1.774	2.626	1.139	0.980	2.080	1.874
	CNN	0.807	0.623	1.004	1.087	1.049	0.875	1.036	0.887
	Transformer	0.653	0.646	0.727	0.778	0.777	0.776	0.420	0.346
IG	LSTM	0.877	0.822	1.696	2.446	1.024	0.960	1.051	0.949
	CNN	0.879	0.643	1.009	1.074	1.152	0.936	1.000	0.847
	Transformer	0.705	0.720	0.798	0.792	0.931	0.817	0.401	0.326
IP	LSTM	1.646	3.354	2.921	2.833	—	—	—	—
	CNN	2.953	3.361	3.347	2.753	—	—	—	—
	Transformer	3.867	3.237	2.937	2.459	—	—	—	—
Attn	LSTM	1.627	3.698	2.555	2.176	2.734	2.544	0.383	0.250
	CNN	3.033	2.204	2.015	1.715	2.699	3.281	0.436	0.234
	Transformer	1.673	1.274	3.287	2.455	0.292	0.305	0.399	0.321

TABLE 3.6: Average SD values over successful and failed test instances.
Colored cells means Success<Fail with statistical significance at a significance level .05 by the unpaired permutation test.

3.7 Results for RQ2: How does the model saliency conformity impact model prediction?

3.7.1 SD scores and Model performance

To investigate the relationship between SD scores and model performance on the task, Sood et al. (2020b) calculated the rank correlation between SD scores and the number of correct predictions in the multiple runs for the same test instance with different random seeds. However, the number of correct predictions in the multiple runs with different seeds measures the model's stability rather than its performance. In addition, they averaged SD scores over multiple runs to calculate the correlation. Therefore, it does not capture a direct relationship between a single SD score and the task result.

Macro-level To remedy the problem, I take a different approach to investigate the SD score and performance relationship. To investigate the overall tendency, the test instances is divided into two classes: correctly-answered successful instances and wrongly-answered failed instances by the model. Then, the average SD scores over each class are compared. Table 3.6 shows the average SD scores for the successful and failed classes for every task. The coloured cells denote that the average SD score of successful test instances is smaller than that of failed test instances with statistical significance at a significance level of .05 by the unpaired permutation test.

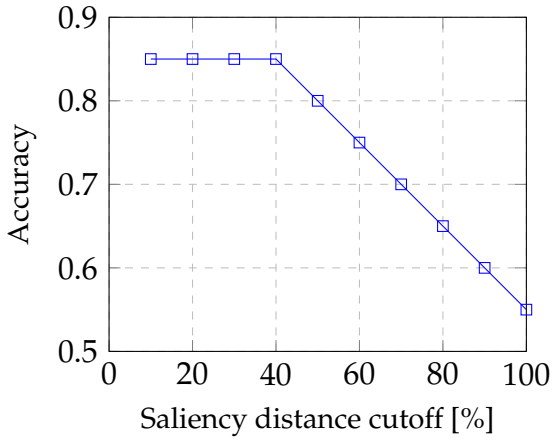
LSTM shows smaller SD scores for successful cases than those of failed cases in both SA and RC tasks but with different interpretation methods, i.e., the IP and Attn methods for the SA task and the gradient-based methods for the RC task. Although other interpretation method and architecture combinations also show smaller successful SD scores than the failed SD scores,

e.g. Grad-CNN and Grad-Transformer for the RC task, their differences are small.

Concerning the QA tasks, a significant result is observed only with the Attn method with CNN and Transformer in the QA-Span task. There is no case for the QA-MC task showing a smaller successful SD score.

Cutoff%	Acc%
10	0.85
20	0.85
30	0.85
40	0.85
50	0.80
60	0.75
70	0.70
80	0.65
90	0.60
100	0.55

(A) Table of binning cutoff from sorted SD scores and cummulated accuracy.



(B) Ideal SDPC Curve for for RQ2 positive answer.

FIGURE 3.9: Ideal model and eye-gaze Saliency distance (SD) with cutoff performance plot.

Micro-level For microscopic investigation, I propose an textbfSD-performance curve (SDPC) to represent the relationship between SD scores and task performance. To draw an SDPC, I first make a list of the test instance predictions arranged by their SD score in ascending order. If RQ2 is affirmative, a sequence of correct predictions followed by wrong predictions is expected. Second, the cumulative model performance is calculated at each increasing 10% cutoff of the prediction list. A cutoff is a group of binned instances from the dataset. After sorting by SD scores, I split the dataset into ten bins. We can decide the number of cutoffs depending on the size of the test set. A smaller cutoff increases the number of bins, making the SDPC smoother. I used a ten per cent cutoff, which means there are ten bins. For example, there are only 45 instances on ZuCo sentiment analysis with gaze and participant answers, so each bin includes four or five test instances. A monotonically decreasing SDPC indicates a positive answer to RQ2, while an increasing curve represents a negative answer to the RQ2. Thus, there are low SD with incorrect prediction instances followed by higher SD with correct prediction instances. Figure 3.9 illustrates the ideal plot for the positive answer to the RQ2. The hypothetical model overall accuracy in this plot is 55%, and the model accuracy start to drop at 50% data. Therefore, SDPC can be a suitable indicator for analysing the relationship between saliency conformity and task performance.

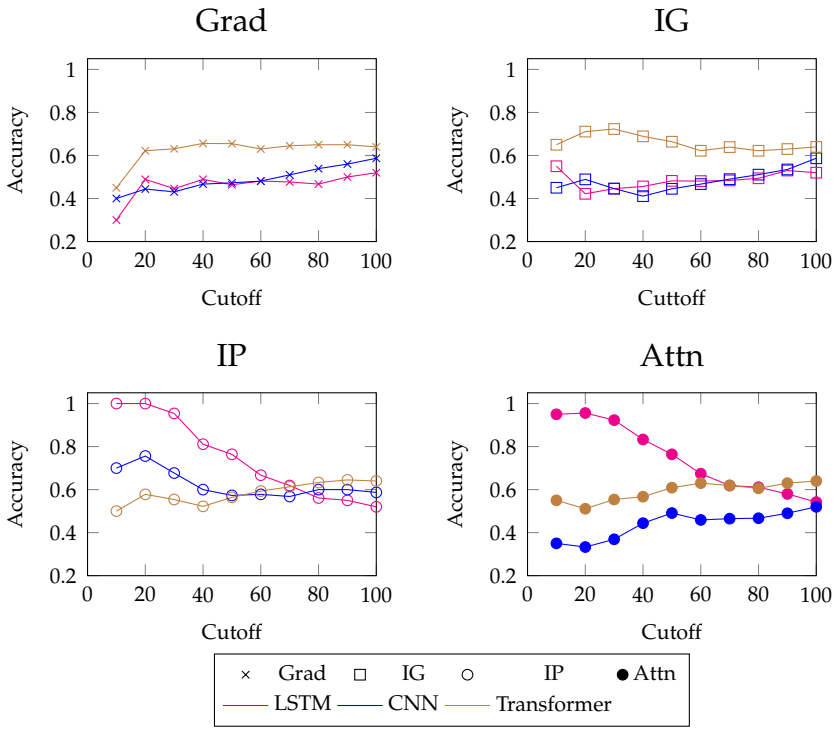


FIGURE 3.10: SD-Performace curves for sentiment analysis (SA)

Sentiment Analysis (SA)

Figure 3.10 shows SDPCs for the combinations of the interpretation methods and the model architectures. The accuracy is used at each cutoff point as a performance indicator for the SA task. Considering the discussion of RQ1, FFD is used as the eye-gaze feature for calculating the SD scores throughout the discussion for RQ2. The accuracies of LSTM, CNN and Transformer on all data (at the right endpoints) are 0.520, 0.587 and 0.640, respectively. The attention layer is introduced to LSTM and CNN to implement the Attn method, causing the different accuracies on all data for the Attn method (the right bottom graph). As a result, the attention layer improves the LSTM performance to 0.542 while degrading the CNN performance to 0.520.

The combinations showing decreasing curves are the IP method with LSTM and CNN, and the Attn method with LSTM. All other combinations show increasing or flat curves. This result agrees with Table 3.6. As discussed for RQ1, Transformer showed good overall conformity with eye-gaze features for the SA task regardless of the interpretation method. However, the SDPC revealed that the overall conformity (the average SD score) does not explain the Transformer's performance. This suggests that focal points contributing to reaching correct answers are different between humans and Transformer. In this respect, LSTM with the IP and Attn method focuses on input words similarly to humans to answer correctly. From the viewpoint of the interpretation methods, the gradient-based methods (Grad and IG) are less suitable for explaining the SA task performance.

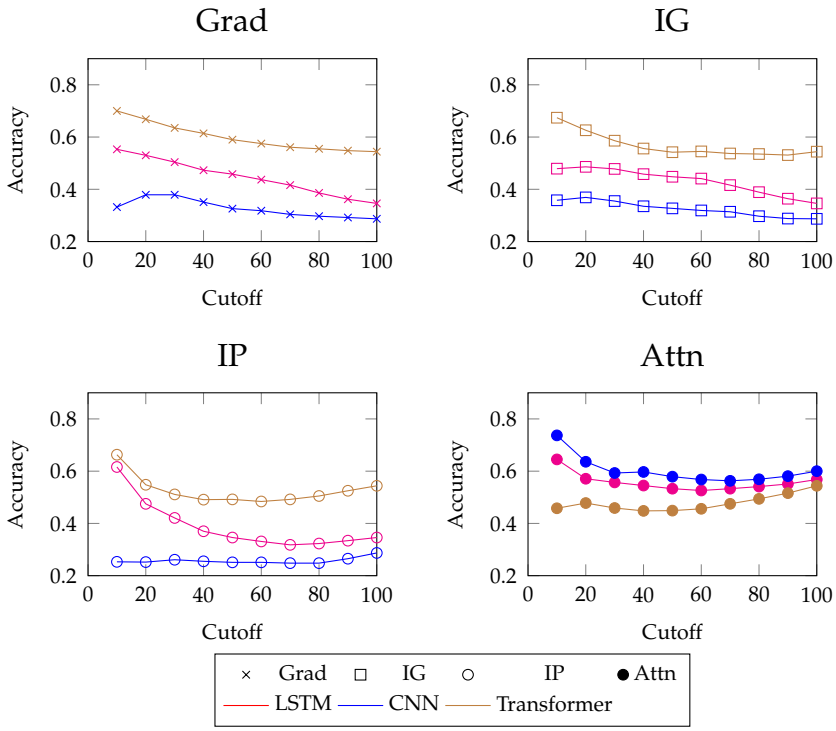


FIGURE 3.11: SD-Performace curves for relation classification (RC)

Relation Classification (RC)

Figure 3.11 shows SDPCs for the combinations of the interpretation methods and the model architectures. I calculate accuracy at each cutoff point as a performance indicator for the RC task. Transformer achieves the overall highest accuracy, 0.544, followed by LSTM (0.346) and CNN (0.287). Adding the attention layer to LSTM and CNN increased the overall performance to 0.568 and 0.600.

The results have more combinations showing the decreasing curves than the SA task. LSTM particularly shows consistent decreasing trends for all interpretation methods. However, this was not captured by Table 3.6 where there is no significant differences in average SD scores between the successful and failed cases for LSTM with the IP and Attn methods. The increasing trend at the later-half cutoff points of their SDPCs might explain this discrepancy. Unlike the SA task, Transformer tends to show decreasing curves except for the Attn method. These trends cannot be seen from Table 3.6. CNN generally shows flat curves. From the viewpoint of the interpretation method, the gradient-based methods (Grad and IG) well explain the good performance of the RC task in contrast with the SA task.

Span-based QA (QA-Span)

Figure 3.12 shows SDPCs for the combinations of the interpretation methods and the model architectures. I use the exact-match (EM) ratio (Rajpurkar et al., 2016) at each cutoff point as a performance indicator for the span-based QA task. Among the nine SDPCs, only the curves of CNN and Transformer

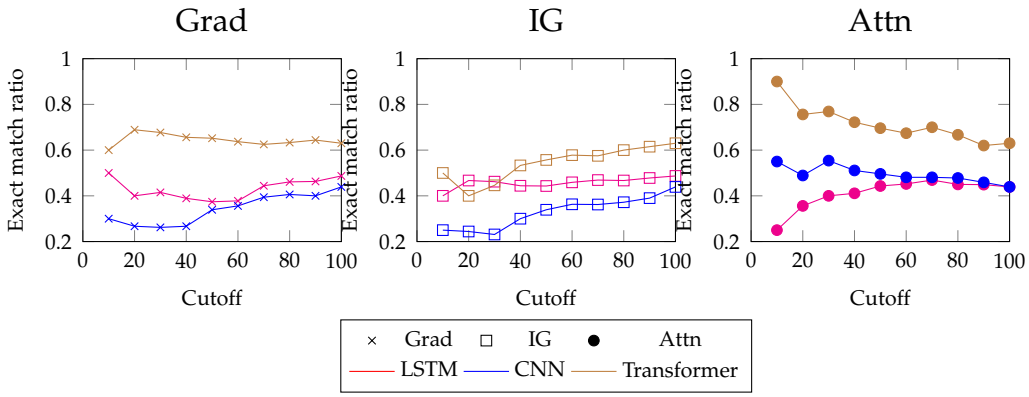


FIGURE 3.12: SD-Performance curves for span-based QA

with the Attn method show a decreasing trend. This result agree with Table 3.6.

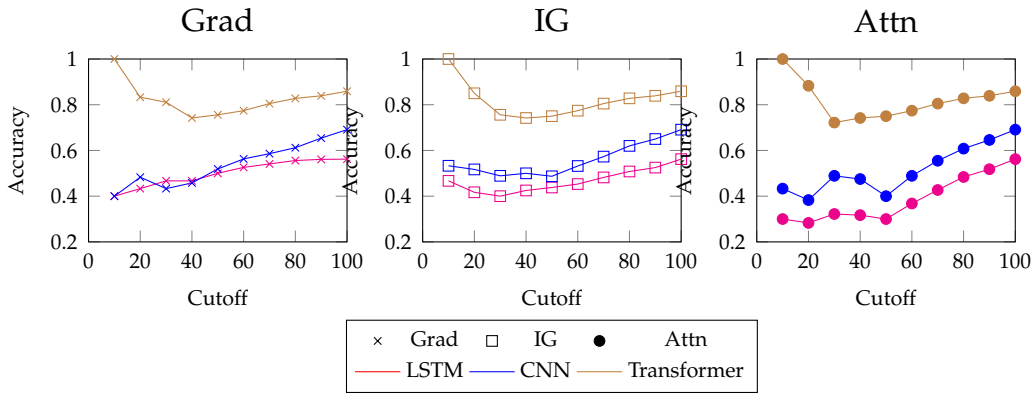


FIGURE 3.13: SD-Performance curves for multiple-choice QA

Multiple-choice QA (QA-MC)

Figure 3.13 shows SDPCs for the combinations of the interpretation methods and the model architectures. In the QA-MC task, accuracy is used at each cutoff point as a performance indicator. Only Transformer shows initial decreasing SDPCs regardless of the interpretation method, but they gradually increase around 30–40 % cutoff. At the 30% cutoff, Transformer performance is lower than total accuracy. Therefore, the only possible curve direction is increasing. This shows that the Transformer answer to the RQ2 are success in the top 30% of instances according to SD scores. The success is defined as low SD with correct prediction followed by higher SD with incorrect prediction. On the other hand, the rest of the instances show fail instances which are high SD with correct prediction answers.

I manually inspected the top 10% and bottom 10% of the data. The general tendency is that model saliency is dense while the human gaze is sparse in the bottom 10%. This suggests that the models' saliency is still high on

What does Ilsa do when Rick refuses to give her the letters?
1: She threatens him with a gun.
2: She throws a stone at him.
3: She tries to persuade him.
4: She threatens him with a knife. 5: She quickly runs away.

Transformer Prediction: 1

Laszlo orders the house band to defiantly play "La Marseillaise". When the band looks to Rick, he nods his head. Laszlo starts singing, alone at first, then patriotic fervor grips the crowd and everyone joins in, drowning out the Germans. In retaliation, Strasser has Renault close the club. That night, Ilsa confronts Rick in the deserted caf. When he refuses to give her the letters, she threatens him with a gun, but then confesses that she still loves him. She explains that when they first met and fell in love in Paris in 1940, she believed that her husband had been killed attempting to escape from a concentration camp. Later, while preparing to flee with Rick from the imminent fall of the city to the German army, she learned that Laszlo was alive and in hiding. She left Rick without explanation to tend her ill husband. Rick's bitterness dissolves. He agrees to help, leading her to believe that she will stay with him when Laszlo leaves. When Laszlo unexpectedly shows up, having narrowly escaped a police raid on a Resistance meeting, Rick has waiter Carl spirit Ilsa away.

Human Answer: 1

Laszlo orders the house band to defiantly play "La Marseillaise". When the band looks to Rick, he nods his head. Laszlo starts singing, alone at first, then patriotic fervor grips the crowd and everyone joins in, drowning out the Germans. In retaliation, Strasser has Renault close the club. That night, Ilsa confronts Rick in the deserted caf. When he refuses to give her the letters, she threatens him with a gun, but then confesses that she still loves him. She explains that when they first met and fell in love in Paris in 1940, she believed that her husband had been killed attempting to escape from a concentration camp. Later, while preparing to flee with Rick from the imminent fall of the city to the German army, she learned that Laszlo was alive and in hiding. She left Rick without explanation to tend her ill husband. Rick's bitterness dissolves. He agrees to help, leading her to believe that she will stay with him when Laszlo leaves. When Laszlo unexpectedly shows up, having narrowly escaped a police raid on a Resistance meeting, Rick has waiter Carl spirit Ilsa away.

FIGURE 3.14: High SD scores and correct predictions in QA-MC task of Transformer with Attn methods.

Interpretation method	Architecture	SA	RC	QA-Span	QA-MC
Grad	LSTM	0.963	2.343	1.057	1.969
	CNN	0.735	1.061	0.951	0.987
	Transformer	**0.656**	**0.752**	**0.778**	**0.410**
IG	LSTM	0.848	2.195	0.991	1.012
	CNN	0.776	1.054	1.028	0.952
	Transformer	**0.699**	**0.794**	**0.887**	**0.390**
IP	LSTM	**2.459**	2.855	—	—
	CNN	3.137	2.921	—	—
	Transformer	3.686	**2.728**	—	—
Attn	LSTM	2.584	2.389	2.635	**0.328**
	CNN	2.648	**1.895**	3.018	0.373
	Transformer	**1.484**	2.901	**0.297**	0.387

TABLE 3.7: Summary of SD scores using FFD and their relation with performance.

the important tokens that are look at by humans, although the entire distributions are different. Figure 3.14 shows the negative case might be happened due to eye-gaze FFD feature is sparse at that instances compare to the top-k instances. The figure shows a question and the multiple choices with Transformer-Attn saliency and human FC along with their answers.

Table 3.6 does not capture these trends at all. LSTM and CNN show increasing trends with all three interpretation methods, resulting in the larger average successful SDs than the average failed SDs in Table 3.6.

3.7.2 Discussion for RQ2

Table 3.7 summarises the average SD scores using FFD excerpted from Table 3.4. The bold number is the lowest SD score for the specific task and interpretation method. The underscored number is the lowest SD score for the specific task and architecture. The coloured cells are overlayed with the significant cells in Table 3.6; in addition, the light-coloured cells denote decreasing trends of SDPCs according to Figure 3.10 through Figure 3.13.

For the SA and RC tasks, The gradient-based methods generally show small average SD scores regardless of the architectures. The small average SD scores can be a good indicator for the RC task performance but not the SA task performance. Even though the IP and Attn methods show large SD scores for the SA task, they explain the task performance well with LSTM. For the QA-Span task, the Attn method with Transformer shows the smallest SD score, and they conform with the task performance. The Attn method with CNN also explains the task performance well despite a large SD score. There is no decisive results for the QA-MC task. However, I observe that Transformer shows small SD scores and a decreasing trend of SDPCs across the interpretation methods.

These results are in line with the study by Feng et al. (2018), which reported that high IG saliency values and small SD scores did not explain the model performance well. They conducted an ablation study of input words according to the IG saliency of the words. Their study revealed that the useful words used to solve the task for humans and those for machines are exclusive to each other. Based on this result, they proposed to use words with high IG saliency as "negative" samples for the entropy regularisation, maximising the model uncertainty. Their ad-hoc proposal might work for a particular task like the SA task, but not for any task; the RC task is a counter-example.

Sood et al. (2020b) proposed a rank correlation between SD scores and the number of correct predictions in the multiple runs with different random seeds. I argued that the number of correct predictions in multiple runs is interpreted as the model's stability rather than its performance. Also, SD scores might be different using different random seeds. Their rank correlation does not directly capture the relationship between SD scores and the model performance. In contrast, the proposed SDPC directly capture this relation. The SDPC enables us to investigate detailed tendencies that are difficult to capture through a single scalar value such as the ranking correlation by Sood et al. (2020b). Without using eye-gaze features directly, we can utilise instances with correct predictions and low conformity. We can use such instances to optimise the models.

To sum up, to what degree does the similarity between the saliency from the interpretation method and human visual attention explain the task performance depends on the task type and model architecture. As far as the experiments I conducted, the Attn interpretation method is robust against differences in tasks and model architectures for explaining task performance. Also, the macroscopic metric, i.e. the average SD score, does not necessarily explain the task performance. The SDPC has limitation when the models attends more tokens, while the humans more focused on smaller number of words thus gaze distribution is more sparse. The models' saliency is still high on the important tokens that are looked at by humans, although the entire distributions are different.

3.8 Chapter Summary

This chapter investigated the interpretation methods for deep neural networks in comparison with human eye-movement behaviour across various NLP tasks. I considered three types of NLP tasks: sentiment analysis (SA), relation classification (RC) and question answering (QA-Span and QA-MC). I also considered four interpretation methods based on simple gradient (Grad), integrated gradient (IG), input-perturbation (IP) and attention (Attn), and three architectures: LSTM, CNN and Transformer. To compare the saliency provided by these interpretation methods and human eye movement behaviour, I used two corpora: the ZuCo dataset and the MQA-RC dataset. Both corpora were annotated with eye-gaze information while humans solved the NLP tasks. I set up the following two research questions. RQ1: *Does the*

	SA	RC	QA-Span	QA-MC
RQ1: Does the input word saliency from interpretation methods conform with human eye gaze features?	Variable Best: Grad+Trans. Worst: IP+Trans.	Variable Best: Grad+Trans. Worst: IP+CNN	Variable Best: Attn+Trans. Worst: Attn+CNN	Variable Best: Attn+LSTM Worst: Grad+LSTM
RQ2: How does the model saliency conformity impact model prediction?	Macro: IP+LSTM IP+Attn Micro: IP+CNN†	Macro: Grad+$*$ IG+LSTM, IG+CNN Micro: IP+LSTM† Attn+LSTM†, Attn+CNN†	Macro: Attn+CNN Micro: Attn+Trans.†	Macro: None Micro: Attn+Trans.†

$*$ denotes any architecture, † denotes positive cases only through SDPC.

TABLE 3.8: Answers to the research questions

input word saliency from interpretation methods conform with human eye gaze features? RQ2: *How does the model saliency conformity impact model prediction?* Through these research questions, I provided the first broad overview of the relation between different interpretation methods and human eye-movement behaviour across different architectures and tasks, which is the first contribution. Table 3.8 summarises the answers to the research questions in this chapter.

To discuss RQ1, I introduced the saliency distance (SD), defined as the KL-divergence between the saliency distributions over input words by an interpretation method and an eye-gaze feature; a small SD score indicates good conformity between human visual attention and focal words by machines. The analysis with SDs revealed that the degree of conformity varied depending on the combinations of the tasks, interpretation methods and architectures, and each question instance. Table 3.8 (the upper half) shows the best and worst combinations of saliency conformity. More concretely, I found that Transformer generally showed small SD scores regardless of the tasks; the gradient-based interpretation methods showed small SD scores for the tasks with relatively short input sentences (SA, RC and QA-Span). I further investigated the saliency distributions over PoS for the SA and RC tasks and those over inside and outside answer spans for the QA tasks. This detailed analysis revealed more precise differences in the saliency distributions between humans and machines, which was not apparent in the overall SD scores.

Concerning RQ2, I investigated the difference between average SD scores over successful and fail test instances to find that a small SD score did not always lead to a correct answer. For detailed analysis, I proposed the SD-performance curve (SDPC) representing cumulative model performance against the SD scores. SDPC enables us to find underlying phenomena that were overlooked using only macroscopic metrics, such as average SD scores and

rank correlations (Sood et al., 2020b). For instance, SDPC told us that Transformer showed a initially decreasing trend of SDPCs for the QA-MC task, even though their average SDs over inside and outside answer spans were not significantly different. The proposal of SDPC constitutes the second contribution of this study.

Overall, good saliency conformity between humans and machines impacts the model performance differently depending on the combinations of tasks, interpretation methods and model architectures. Table 3.8 (the bottom half) summarises the combinations in which good saliency conformity leads to good performance. The results can be divided into macro and micro positive answers. In macro metric, denotes at least one affirmative combination showing the significant difference in average SD scores over successful and failed instances and supported by SDPC analyis. In micro, indicates the the positive answers(†) combinations supported by the SDPC analysis only and the plot might show a negative answer at the other part of the plot.

Chapter 4

Interpreting Models in summarisation

Previous studies (Feng et al., 2018; Hollenstein and Beinborn, 2021b; Ikhwantri et al., 2023; Jawahar, Sagot, and Seddah, 2019; Sood et al., 2020b) have extensively investigated various aspects of interpreting classification models for NLP, such as feature attribution (Poerner, Schütze, and Roth, 2018) to human behaviour such as eye gaze. However, the application to sequence models (Alvarez-Melis and Jaakkola, 2017; Vafa et al., 2021), such as summarisation tasks (Xu and Durrett, 2021), introduces additional complexities that make these techniques challenging to apply directly. As a comparison, the output of interpretation classification models is a single output from a binary, multi-class, or multilabel. Meanwhile, summarisation models make sequential decisions based on an extensive vocabulary. In addition, encoder-decoder models have a different architecture, featuring a complex interaction of decoder-side and encoder-side computation to select the next word.

In this chapter, I proposed a framework to analyse the model behaviour in summarisation by comparing it to human summarisation behaviour using eye-gaze data. For RQ1, I measure conformity by calculating the correlation between encoder-decoder summarisation model saliency and human fixation counts. For RQ2, I conducted ablation experiments by removing parts of inputs such as words, phrases, or sentences that sequence models or humans consider necessary. In the following sections, I will explain the methodology used to interpret the sequence model used in the proposed frameworks and the method to calculate saliency in the sequential encoder-decoder model.

4.1 Interpreting Sequence Model

In the previous chapter, I compared the saliency vector and eye-gaze feature vector to analyse the behaviour of reading tasks. These vectors have the same dimension of input length. A saliency vector correspond to a model prediction. On the other hand, in the writing task, the interpretation method outputs a sequence of saliency vectors, corresponding to the output tokens. This can be viewed as a matrix form (Sequential saliency matrix, Figure 4.1), where the rows represent the input tokens, and the columns represent the output tokens. The eye-gaze feature vector is also used for writing activity in human translation (Carl, 2012a).

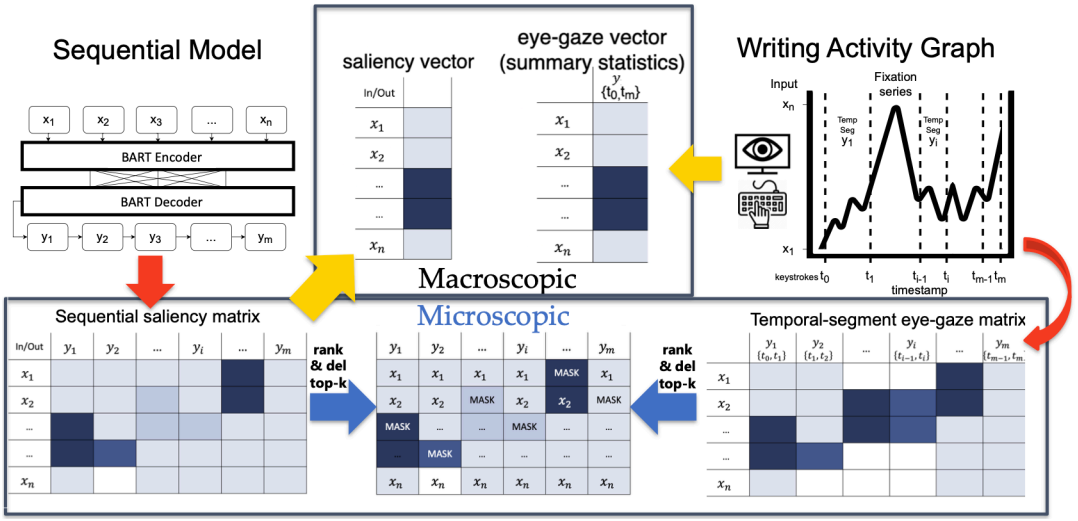


FIGURE 4.1: Proposed framework consists of macroscopic and microscopic analysis between model saliency and human gaze data. The macroscopic analysis compares the model saliency and eye gaze after aggregation across the entire output generation, while the microscopic analysis compares them at each output token.

To compare these different representations of model's saliency matrix $(n \times m)$ and eye-gaze feature vector (n), I proposed two approaches, macroscopic and microscopic analyses. In macroscopic analysis, I analyse the model saliency and eye-gaze relationship by collapsing the saliency matrix into a single vector. The collapsing is achieved by word-wise aggregation of the saliency scores across all outputs, transforming into a saliency vector for the source text in which the dimensions correspond to the word token. In microscopic analysis, I propose a new feature representation for eye gaze from raw temporal data into discrete temporal segments that align into the sequential saliency matrix (Figure 4.1). This feature representation allows analysis of input attribution between the model saliency and eye gaze at each output token. Integrating sequential model saliency outputs with eye-gaze data is a novel approach that presents two primary challenges: the different representations of sequential saliency matrix from models and eye-gaze features at the word level from eye-tracking raw data.

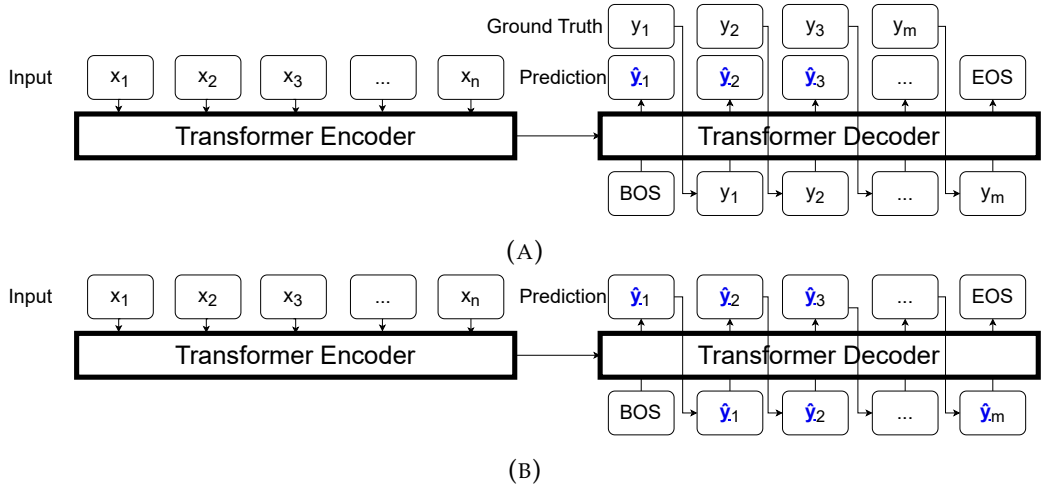


FIGURE 4.2: (A) Teacher forcing at training (B) Auto-regressive at testing

4.2 Generation method

To extract the saliency of a model output, I need the model to produce an output prediction or be given a ground-truth token. In terms of producing output, a classification during testing is simply by calculating an arg max over the output class probability distribution. However, the sequential encoder-decoder model has a different mode of generating output during the training and prediction phase.

During the training phase, the decoder was given the previous ground truth in the generation process to predict the subsequent step output (Figure 4.2(A)). This training phase is called teacher-forcing. During this phase, the model selects the output greedily using the maximum logit probabilities for each step and minimizes it with ground-truth output.

In the prediction phase, the models use their prediction as an input to their decoder instead of ground truth. This difference could make a difference in performance due to ground-truth exposure bias and greedy max use during training. A sequence decoding method such as beam search (Freitag and Al-Onaizan, 2017; Nallapati et al., 2016) is usually applied to decoder output to select the optimal probability of a sequence effectively. The beam search method traverses the most optimal left-to-right path by keeping a fixed amount of candidates. Figure 4.2(B) illustrates the prediction phase called free-decoding in this study.

As the possible combination of output space between machine and human can be very large, I use the force-decoding method from human summary to extract the most optimal saliency matrix in this study. This approach is usually applied for interpreting sequential model (Xu and Durrett, 2021). It guarantees the number of columns of the saliency matrix to be the same length as the human summary. However, it does not necessarily mean the model outputs the exact token prediction when applying beam search or other decoding methods. For the rest of this chapter, I use forced decoding to generate model prediction unless mentioned otherwise.

4.3 Eye-gaze Data for summarisation

I use three eye-gaze datasets collected during human summarisation. Table 4.1 shows the statistics of the datasets. SSG23 is a newly collected eye-gaze data in this study.

Dataset	CS19	IELTS33	SSG23
#Participants	13	11	30
#Source texts	6	3	2
#Summaries	26	33	53
Ave. source length [word]	141	867	463.7
Ave. source length [sentence]	6.3	15	21.5
Ave. reading time [min]	0.5	4	4.9
Ave. writing time [min]	6	14	20.7
Reduction rate [%]	80	22	25

TABLE 4.1: Statistics of eye-gaze datasets.

4.3.1 CS19

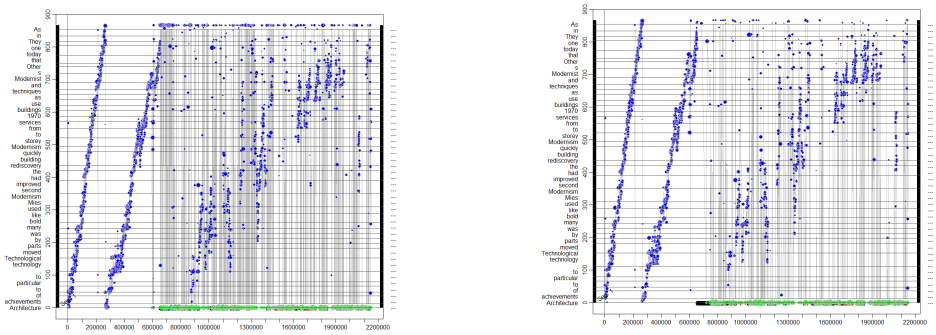
Sahoo and Carl (2019) collected eye-gaze data for three writing tasks: text copying, paraphrasing, and summarisation in English. I use their summarisation data named CS19 in this study. Thirteen people wrote a total of 26 summaries from six source texts. Four of the six source texts are news articles, and the other two are sociological texts from an encyclopedia. At least four people from the participant pool covered each source text.

4.3.2 IELTS33

Dataset statistics The eye-gaze data during summarisation was collected using Translog-II. In the collection process, three texts were randomly selected from the English proficiency test IELTS[1] that would fit on the screen while excluding texts that were too similar. The texts describe scientific discussions on the effects of noise, 20th-century architecture, and endangered languages, which are 834, 955, and 813 words in length, respectively. There are 11 participants, 10 males and one female, for the data collection experiment. The participants are mostly native speakers or near-native speakers of English proficiency who are studying at a UK university. One was a master's-level computer science professional, four were undergraduate students, and seven were PhD students.

The participants used a workstation equipped with an eye tracker (Tobbi Pro X3-120) and Translog-II software running on Windows. The participants use a chinrest, which helps to fix the distance between the eyes and the screen and boosts the eye-tracker's overall accuracy without being overly intrusive. Three observed summarisation sessions were conducted following

[1] https://ielts.org/



(A) Original (B) Corrected

FIGURE 4.3: Vertical error noise and correction from eye-tracking based on Mishra, Carl, and Bhattacharyya (2012)

a brief training phase. Each session began with a calibration phase on Tobii, with brief breaks between sessions. An entire session with three texts takes, on average, 1.5 hours. There are 33 summaries, 11 for each source text from the collected dataset in total.

Post-process After the data collection, I noticed four participants had a high rate of vertical errors in the eye tracker. To remedy the vertical errors, I applied an error correction algorithm (Mishra, Carl, and Bhattacharyya, 2012), which vertically shifts the fixations to the nearest line of the previous gaze and discards the vertical jump based on the past and future gaze location. The algorithm was applied to all data, setting the algorithm threshold so that the correction did not affect the other less erroneous data. To check the validity of the correction, I manually compared the data before and after the correction.

4.3.3 SSG23

Corpus Collection In this study, I construct an eye-tracking corpus of L2 English speakers in writing summaries from L1 Japanese speakers. The dataset was designed to be annotated with both explicit annotations of marking important text spans and implicit information from eye gaze. This eye-tracking corpora is designed to record the process of extracting core information and writing a summary. I collect 53 summaries from two source texts. The source text is developed for assessing the use of argument framework for University-level L2 students (Sawaki, 2020). The source texts describe a scientific report about a water purifier product (Cycoclean) and the effects of short napping on memory (Napping and Learning). There are five expert summaries for the Cycoclean text and six expert summaries for the 'Napping and Learning' source text. I compared the students' summaries to the expert summaries to evaluate the quality of the collected student summaries. The student's summaries were obtained by the corresponding students' eye-tracking logs. The summaries are written by 30 university students (undergraduate and graduate level) recruited from a Japanese university.

Experiment Scenario and Procedure The experiment was designed to capture both implicit information from eye gaze and explicit annotation of marking important text spans, which records the process of writing a summary. To fulfil this requirement, I used Translog-II (Carl, 2012b) eye-tracking tools which have been used in Translation studies (Jakobsen, 1999). However, the program lacks a feature to highlight important text spans, so I added this functionality. The highlighting events of the text span are recorded in the event log data along with other event data, such as fixation and keystroke events.

In the experiments, I used a 23.8-inch LCD monitor with an Infrared eye-tracker Tobii Pro X3-120, which has a sampling rate of 120Hz. In the experiment setup, the eye tracker is connected to an auxiliary computing unit to reduce the main PC's computational burden. I set up the experiments in Translog-II for the full-screen 1920x1080 setup to minimise the possible disruption. In addition, keys other than alphanumeric, symbol keys, return and delete keys were disabled. The modified keyboard's purpose is to only write a textin the experiment. It is also to avoid possible disruption. For example, a window button might pop up a start menu during the experiment. It can cause a distraction to the participant. This accidental distraction during the experiment could render the eye-tracking data invalid and might cause the experiment session to fail.

The participants are instructed to watch a lecture video that explains the summarisation task definition before starting the experiments. The video duration is about 25 minutes. Before the experiments, I also introduce the participant to the tool's interface with a sample English text. The experiments started by calibrating the eye-tracker with the participant. After finishing the calibration phase, the participants are instructed to do a full reading first and highlight important information, which I call this phase reading and extractive summary. After that, the participants continued to write the summary while still having access to the source texts and their highlight information. The participants are encouraged to write a highly abstractive summary of about 80 words in the application interface.

The experiment recruitment for participants was open and conducted for around seven months. The participants were fairly compensated for their participation, with some monetary benefits of 3,000 JPY for 90 minutes in total. This collected dataset is referred to as Student summarisation Gaze (SSG23) in the rest of this thesis.

Post-process The same post-process method applied to the previous dataset (Subsection 4.3.2). I applied an error correction algorithm (Mishra, Carl, and Bhattacharyya, 2012), which vertically shifts the fixations to the nearest line of the previous gaze and discards the vertical jump based on the past and future gaze location. The algorithm was applied to all data, with the algorithm threshold so that the correction did not affect the other less erroneous data. To check the validity of the correction, I manually compared the data before and after the correction.

Text	Nap.& Learn.	Cycoclean	Total
#Summaries	26	27	53
#Token Src	481.0	447.0	463.7
Avg of #Token Sum	116.5	121.1	118.8
Avg of Len Sum	616.2	628.1	622.3
Avg of #FixCount Src	776.3	544.6	658.2
Avg of #FixCount Sum	623.0	673.7	648.8
Avg of Duration (Mins)	25.8	25.7	25.7
Avg of Reading Time Src (Mins)	5.9	4.1	4.9
Avg of Reading Time Sum (Mins)	5.6	6.5	6.1
Avg of Inserted Chars	894.0	910.1	902.2
Avg of Deleted Chars	277.2	278.4	277.8

TABLE 4.2: Statistics of the student summarization dataset (SSG23).

Dataset Statistics Table 4.2 shows the statistics of written summaries, experiment duration, fixations, and reading time. Overall, the participant wrote around 118 words. The average words (**#Token Src**) in each source text are 116 and 121 in Napping and Learning and Cycoclean, respectively. Napping and Learning source text is fixated more than its written summary (**Avg of #FixCount Sum**). On average, in Cycoclean, the source text is less fixated than its written summaries. The average session duration (**Avg of Duration**) is around 25 minutes per text. The average reading time (**Avg of Reading Time**) is the sum of total fixation durations on a particular area of interest, such as a token. The average reading times have a similar trend with the total fixation counts. Inserted chars (**total of Inserted Chars**) show how many characters are typed by participants. In contrast, deleted chars track how many inserted chars are deleted during an experiment session. The deleted chars (**total of Deleted Chars**) show the revision process by a participant, which does not happen in the machine.

Table 4.3 shows the statistics of written summaries per participant. The table column **#text** shows how many summaries a participant finishes. Dur is the duration (in minutes). ∑TrtS and ∑TrtT are the total reading time (TrT) of Source (S) and summary(T), respectively. FixS and FixT are the fixation counts for Source (S) and Summary (T), respectively. ∑#Ins and ∑#Del are the total inserted and deleted characters in the summaries, respectively, which are not only the final version but also include the revision stage. TokT and LenT are the average length of tokens and the length of characters in the final version of summaries. Overall, there are 6,299 tokens in written summaries from the participants. In total, the rate of revision in total is around 30% from total #inserted chars divided by #deleted chars. Source text is fixated more than target text on average per participant.

Part	#Text	Dur	∑TrtS	∑TrtT	FixS	FixT	∑#Ins	∑#Del	TokT	LenT
P01	1	16.5	3.8	2.4	708	430	723	122	112	601
P02	2	29.6	5.4	9.6	382.5	454.5	1434	300	222	1134
P03	1	40.4	5.2	8.4	849	1407	875	362	98	519
P04	2	71.5	13.0	16.1	613.5	637.5	1661	477	214	1096
P05	2	66.0	10.1	14.1	552.5	568.5	1463	693	148	767
P06	2	60.1	9.8	19.7	1033	1759.5	1578	405	224	1173
P07	2	57.5	13.1	14.7	1248.5	1130	2143	713	262	1430
P08	2	61.2	13.3	15.3	1202.5	967	2470	857	315	1612
P09	1	47.6	12.1	9.9	1196	830	1234	560	121	674
P10	2	35.6	7.4	8.5	523.5	395.5	1683	506	223	1175
P11	2	35.9	6.8	17.7	341.5	563	2066	599	290	1467
P12	1	45.8	6.2	20.7	648	1686	651	192	89	458
P13	1	15.7	2.6	1.4	407	157	945	199	137	746
P14	2	42.3	9.1	11.1	543	515	2190	584	301	1606
P15	1	52.1	6.9	21.3	1100	2651	1227	465	143	762
P16	2	29.2	4.7	11.9	327.5	396	1347	424	171	923
P17	2	61.2	6.5	4.0	224	149	1837	357	290	1480
P18	2	76.3	19.1	4.2	1086.5	246.5	2312	871	279	1441
P19	2	29.0	8.1	1.8	521.5	119.5	1295	352	168	943
P20	1	33.1	5.0	7.0	401	473	833	317	100	516
P21	2	41.5	9.5	8.9	587	457.5	1956	347	299	1609
P22	2	27.9	6.2	5.0	419	210.5	1738	300	284	1437
P23	2	53.1	9.5	8.2	521	209	1868	822	195	1042
P24	2	45.2	7.1	8.9	423.5	503	1316	316	196	1000
P25	2	46.0	9.4	13.3	864.5	1109.5	2120	722	264	1387
P26	2	48.7	12.5	8.9	850.5	529.5	1727	662	203	1064
P27	2	39.0	6.2	10.8	589.5	703.5	1187	406	153	781
P28	2	44.7	6.1	14.2	404	766	1481	161	258	1318
P29	2	47.6	12.2	5.1	657	203.5	1780	435	264	1344
P30	2	64.2	14.7	19.7	873	784	2677	1200	276	1477
Total	53	1364.6	261.9	323.0	658.2	648.8	47817	14726	6299	32982

TABLE 4.3: The dataset statistics per-participant (Part). #Text is the number of summary finished in an experiment.

4.4 Experimental Setting

These experiments use BART (Lewis et al., 2020) as the primary target architecture as it is widely used in summarisation task [2]. I use three BART variations: the pre-trained BART model (BART-PT[3]), the finetuned BART model by the XSum dataset (Narayan, Cohen, and Lapata, 2018) (BART-FT)[4], and a distilled BART model (DistilBART[5]) which also finetuned by the XSum dataset. I do not use the dataset including eye-gaze information for finetuning. The reason is that the eye-gaze data is small, and we use all the data for the evaluation study and analysis instead of directly improving the model performance.

I use four interpretation methods: Input Gradient (Grad), Integrated Gradient (IG) (Shrikumar, Greenside, and Kundaje, 2017), Occlusion (Occ) (Zeiler and Fergus, 2014) and Attention (Attn) (Bahdanau, Cho, and Bengio, 2014). Following Xu and Durrett (2021), the attention is obtained from the last layer cross-attention matrix of the BART model. The cross-attention vector over input is obtained by calculating mean over decoder output of the cross-attention matrix. The gradient-based methods obtained the results from the embedding layer of the encoder module as similar to the task-specific reading and the previous study. In addition, I also use two baselines: Random, which randomly assigns a token distinct integer representing an importance ranking, and Lead, which assigns a token a rank based on its input position.

As an eye-gaze feature, I use fixation count (FC), defined as the number of fixations on a specified object during a specified duration. The fixation count is used because the features capture all processing stages from the start of the reading to the writing of the final character. Although I also used the first fixation duration (FFD) in the previous chapter (Chapter 3), it only represent the first fixation duration on words and captures early reading stages. It does not capture the human fixation on writing phase.

To extract the saliency matrix, I run the summarisation models using force decoding to generate human summaries. In force decoding, the models predict the next tokens given the previous ground-truth tokens. This does not mean the models will make a correct prediction, but the models will have an upper bound of the conditional probability distribution similar to that of the training mode.

4.5 summarisation Evaluation

I evaluate the models using the participants' written summaries to perceive the realistic performance in our experiment datasets before extracting the model's saliency in the next section. Table 4.4 shows the models' performance without force decoding.

[2] Based on the Huggingface download metrics search Link on 13th Oct 2023
[3] https://huggingface.co/facebook/bart-large
[4] https://huggingface.co/facebook/bart-large-xsum
[5] https://huggingface.co/sshleifer/distilbart-xsum-6-6

Model		BART-PT			BART-FT			DistilBART		
Dataset	TextID	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
CS19	U1	.399	.158	.244	.243	.036	.154	.259	.030	.162
	U2	.279	.075	.165	.249	.057	.158	.218	.024	.150
	U3	.518	.239	.318	.429	.146	.266	.400	.164	.224
	U4	.507	.245	.273	.380	.148	.227	.375	.088	.229
	U5	.539	.230	.383	.284	.043	.193	.328	.093	.214
	U6	.440	.137	.235	.230	.056	.159	.244	.040	.168
	mean	.447	.180	.269	.303	.081	.193	.304	.073	.191
	std	.090	.063	.068	.075	.047	.042	.068	.049	.032
IELTS33	U1	.417	.117	.204	.257	.051	.167	.245	.035	.141
	U2	.349	.068	.165	.322	.064	.178	.291	.050	.170
	U3	.419	.129	.205	.281	.056	.142	.266	.046	.147
	mean	.395	.105	.191	.287	.057	.162	.267	.044	.153
	std	.033	.026	.019	.027	.005	.015	.019	.006	.013
SSG23	U1	.301	.061	.151	.303	.047	.149	.244	.029	.141
	U2	.418	.138	.224	.246	.042	.147	.323	.062	.182
	mean	.360	.100	.187	.274	.045	.148	.283	.045	.161
	std	.058	.038	.036	.028	.002	.001	.039	.017	.020

TABLE 4.4: Rouge scores of Model's summaries against participant's summarirs.

CS19 From the table, the pre-trained BART model (BART-PT) shows high performance on the CS19 dataset compared to the finetuned model (BART-FT) and distilled model (DistilBART). BART-PT obtained 0.447 Rouge-1 and 0.18 Rouge-2. Source text U3 on CS19 appears to be the maximum Rouge score obtained by the summarisation models. Compared to other datasets, CS19 have the highest Rouge score obtained by the summarisation models in this study. This result is expected, given the CS19 source texts are obtained from the news dataset.

IELTS33 Similar to the CS19 dataset, the BART-PT model obtained the highest performance in the dataset. However, the performance gap between BART-PT and other models is quite large. The closest gap is 0.1 in Rouge-1, 0.05 in Rouge-2, and 0.02 in Rouge-L from BART-FT. This is unexpected because finetuned models should perform better in their own optimized tasks. However, there are several possibilities, such as task-specific issues or distribution mismatch. Specifically, I hypothesise that IELTS33 text distribution is out-of-domain of the finetuned dataset, which leads to a poor generalisation (Kumar et al., 2022). This is because the IELTS33 text has an academic style of writing, while BART-FT pre-trained on the XSum dataset has a news-article style that puts important information in the early part of the text.

SSG23 Overall, the average performance in SSG23 dataset is also the same where BART-PT have higher Rouge-score than BART-FT. There is one noticeable result where BART-FT Rouge-1 is slightly higher than BART-PT on U1 text instance. However, the other Rouge-2 and Rouge-L metric show that

BART-PT still produce better summary when evaluate to all 30 students. Similar to IELTS33 result, the BART-FT overall performance is higher than DistilBART. This reason might be because of the same academic report domain of the IELTS33 and the SSG23 datasets.

4.6 Macroscopic Analysis

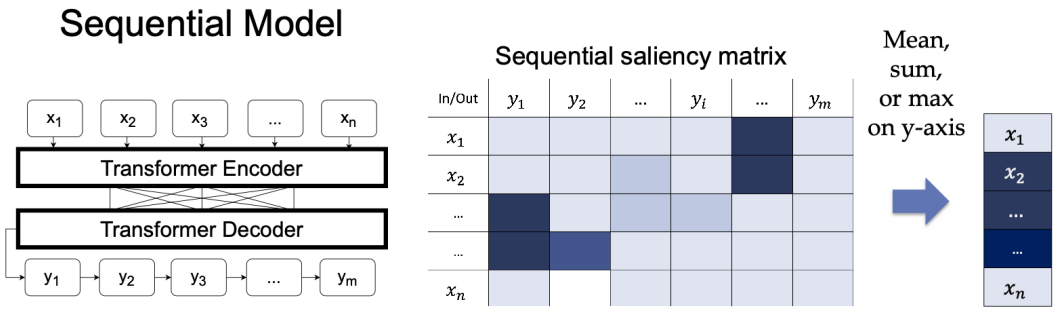


FIGURE 4.4: Aggregation of model saliency output for each sequence step.

This section analyses the correlation between the model saliency and fixation counts over the summary generation process. I calculate the saliency vector by aggregating the columns of a sequential saliency matrix.

Token-level (token) For each source text, a word-wise aggregation of the saliency scores across the entire generation steps is conducted. As a result, a saliency vector for the source text is obtained (Figure 4.4). The dimensions of the saliency vector correspond to the word token, and their values indicate the word saliency. I consider two aggregation methods, max and mean. The max aggregation takes the maximum saliency score across all word generation steps as a saliency vector element, while the mean aggregation takes the average saliency scores. Likewise, I aggregate the fixation counts of each word token in the source text by summing up them across the summary writing process to obtain a fixation count (FC) vector for the source text.

I consider a written summary from the start at the first character input of the summary. In the following experiments, I do not distinguish fixations from different participants in the summation to calculate the correlation with summarisation models.

Segment-level (seg.) A word might be too small as a unit. At the same time, the sentence is a fixed boundary that is too large as a linguistic boundary to analyse the relationship between the model saliency and fixation counts. Some larger linguistic units than words, such as phrases and clauses, should have been considered for aggregation (Kamp, Beinborn, and Fokkens, 2024). For this reason, I also consider a discourse segment-wise (Marcu, 2000; Marcu

Sentence 1	It is difficult to be settled under such a subsistence system as the resources of one region can quickly become exhausted.
Sentence 2	Hunter-gatherer societies also tend to have very low population densities as a result of their subsistence system.

FIGURE 4.5: EDU-segment inside the sentence colored in different segments.

and Echihabi, 2002) saliency vector. I adopt the Elementary Discourse Unit (EDU) (Marcu, 2000) for segmentation. EDU segmentation is a clause based segmentation. It parses text into segment from patterns of usage of cue phrases such as "however" and "in addition to". Previous studies has demonstrate that discourse segments can be used to improve summarization models in supervised (Xu et al., 2020) or unsupervised (Dong, Mircea, and Cheung, 2021) manners. The vector dimension corresponds to an EDU segment. I use SegBot (Li, Sun, and Joty, 2018) to segment the text and map the word into segment features. The vector values are calculated by averaging the token values in the segment. Figure 4.5 shows the EDU segments compared to the tokens and the sentence. It shows three EDU segments inside a sentence. These segments provide more meaningful information than a single token but are more fine-grained than a sentence.

Sentence-level (sent.) I also consider a sentence-wise saliency vector and an FC vector in which the vector dimension corresponds to a sentence in the source text, and its value denotes the sentence saliency or sentence fixation counts. The vector values are calculated by averaging the token values in the sentence.

4.6.1 Discussion for RQ1

To answer RQ1, I calculate Spearman's rank correlation ρ between the saliency vector and FC vector of each source text. For each source text correlation, the average value of Spearman's ρ is computed across all source texts by transforming the value into the Fisher's z score and converting it back to the ρ value (Myers and Sirois, 2006). I use rank correlation instead of KL-divergence (Chapter 3) because The KL-divergence method might not be robust when the model saliency distribution contains zero probability. In the preliminary experiments of the summarisation task, I observed the many inf and -inf results because many tokens have zero saliencies. Thus, I opted to use rank correlation, which we can interpret similarly to KL-divergence according to the previous study (Hollenstein and Beinborn, 2021b).

Table 4.5 shows the correlation between the model saliency and fixation counts averaged over the source texts. Overall, the max aggregation tends to provide higher correlations than the mean aggregation except for Occ. The mean aggregation normalises the total saliency score by the output summary length. Therefore, it pushes down the aggregated saliency score of words that are highly salient in a few word generation steps. It can be concluded that the

Dataset		CS19			IELTS33			SSG23		
Model+Interp.	Aggr.	token	seg	sent.	token	seg	sent.	token	seg	sent.
Random		.114	-.004	.292	.052	.014	.143	.001	.000	.009
Lead		-.541	.543	.762	-.319	.236	.012	-.269	.173	.392
BART-PT Grad	max	.360	-.172	.307	.401	-.147	-.088	.211	-.192	-.043
	mean	-.092	-.204	-.042	.003	-.152	-.066	.120	-.232	-.187
BART-PT IG	max	.299	-.018	-.007	.329	-.018	.131	.202	-.281	-.181
	mean	-.018	-.218	-.200	.012	-.090	.140	.130	-.281	-.234
BART-PT Occ	max	.009	.566	-.380	-.046	.089	-.317	-.014	-.008	.019
	mean	.059	.266	.203	.014	.013	-.202	.046	-.111	-.190
BART-PT Attn	max	.420	.254	.645	.338	-.018	.080	.385	-.185	-.141
	mean	.477	-.098	.455	.289	.111	.297	.365	-.163	-.111
BART-FT Grad	max	.441	.111	.282	.269	-.069	-.010	.145	-.158	-.023
	mean	.227	-.165	.207	.050	-.182	-.127	.069	-.237	-.098
BART-FT IG	max	.433	.165	.644	.222	-.029	-.087	.181	-.301	-.098
	mean	.238	-.224	-.002	.064	-.186	.044	.118	-.421	-.015
BART-FT Occ	max	.219	-.241	-.360	.076	.067	.236	.067	.076	-.029
	mean	-.029	.050	.129	.102	.026	.149	.033	-.022	-.204
BART-FT Attn	max	.545	.098	.251	.395	.044	.058	.353	-.190	.004
	mean	.488	-.026	.659	.281	.104	.004	.301	-.082	.176
DistilBART Grad	max	.319	.161	.362	.222	-.108	-.105	.194	-.158	-.011
	mean	.089	-.213	.172	-.009	-.233	-.391	.132	-.174	-.034
DistilBART IG	max	.399	.287	.204	.257	-.204	.108	.208	-.219	-.048
	mean	.113	-.118	.045	.032	-.245	-.057	.128	-.230	-.026
DistilBART Occ	max	.002	-.233	.513	-.049	-.119	NaN	.039	.107	.072
	mean	.117	.143	-.363	.088	.110	.400	.074	-.032	-.012
DistilBART Attn	max	.510	.296	.546	.401	-.051	.020	.389	-.098	-.058
	mean	.457	-.155	.423	.263	.046	-.129	.337	.050	.109

TABLE 4.5: Average rank correlation between model saliency and fixation counts.

mean aggregation is inappropriate for the macroscopic analysis. I focus the result by the max aggregation (coloured rows) in the following discussion.

Among the interpretation methods, Attn shows stable high correlations, particularly on token-based correlations, followed by the gradient-based methods (Grad and IG). The Lead method shows a significantly high sentence-based correlation for CS19. This high correlation can be explained by domain bias; four out of six source texts in CS19 are news articles. In the news domain, important information tends to be placed in the earlier part of texts due to the writing convention in journalism. This explanation is supported by the Lead model's low token-based correlation in CS19 and the low correlation in IELTS33. The source texts of IELTS33 are scientific articles, which have a different writing style from news articles.

Comparing the first two datasets, the correlations in CS19 tend to be higher than those of IELTS33. The difference in source text length, 141 vs 867 words on average, can explain this tendency. In addition, the reduction rate of IELTS33 is four times higher than that of CS19. In CS19, the reduction rate is 80%, which means summarisation removes about one sentence from the source text. I expect that the source text and summaries will be very similar, leading to high correlations in both token and sentence-based metrics. On the contrary, the IELTS33 texts are lengthy, and the reduction rate is high, i.e., summaries are one-fourth of the source texts in length. I expect more various operations, such as paraphrasing, splitting sentences, and deleting sentences, applied in the IELTS33 summaries, which makes the alignment between the source text and summaries more difficult. The low correlations of IELTS33, particularly in the sentence-based metric, support this explanation. I also notice the difference in the token-based correlation between the datasets is larger in the finetuned models (BART-FT and DistilBART) than in the pre-trained model (BART-PT). This difference can also be explained by domain bias, which has been mentioned above. Since the news articles are dominant in the XSum dataset, the former models successfully finetuned which words in the source text to focus on when summarising a news article, which is also dominant in CS19.

In SSG23, the correlations at the segment level tend to be negative. At the same time, the Lead shows a weak positive correlation at the segment level and more so at the sentence level. This negative correlation means the human gaze and machine saliency prioritize different directions. The negative correlation between humans might indicate a wrong direction for the machine. However, the negative correlation between machines and humans cements the importance of integrating human cognition into machines, which is shown by Feng et al. (2018). Another possible reason is the tokenization method. Specifically, aggregating saliency of the BART tokenization methods affects longer subwords within the segment. Thus, using a method that redistributes model tokens distribution (Voita, Sennrich, and Titov, 2021) might be better.

Regarding the models, I observe consistently high token-based correlations by the BART-FT model in CS19, regardless of the interpretation method. However, I can not observe clear differences among models in the other

columns in Table 4.5.

The answer to RQ1 is that I observe weak correlations between word saliency from the interpretation method and fixation counts on words under some conditions. First, the models show weak to moderate correlations, mostly within the same text domain of the modes trained. This means the models might learn the domain bias of the news summary patterns instead of learning to select important information. Second, the models' saliencies are better correlated at higher levels of input, such as segments or sentences within the same domains such as CS19 dataset. However, outside the model trained domain, the models' saliencies are better correlated at token input. Last, the attention-based interpretation method (Attn) is promising due to its higher correlation than any other method at the macroscopic level. This result is similar to our findings on question-answering tasks (Chapter 3), which have almost the same complexity as the summarisation task.

4.6.2 Discussion for RQ2

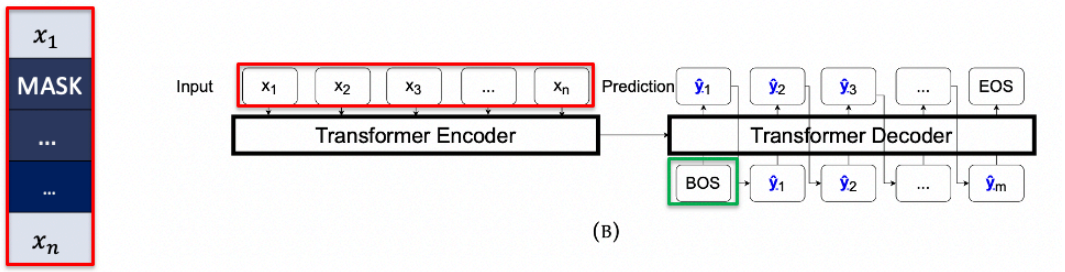


FIGURE 4.6: Macroscopic input ablation using free-decoding

To answer RQ2 at macroscopic levels, I propose an input ablation based on an aggregated saliency matrix. The models predict the next token using free-decoding methods. Figure 4.6 illustrates the input to the models (highlighted in red). I apply an input ablation by removing top-k salient input at different levels, such as tokens, EDU segments, and sentences. I use the beam search method and apply several post-processing methods, such as the length-constrained output. I set the length constraint relative to the target output. The minimum range is set to the length of target output minus 5, and the maximum range to the length of target output plus 5. I also prevent the outputs from repeating the same bi-gram [6] which similar with Keskar et al. (2019). I evaluate the generated summary from the free-decoding output using the Rouge score.

First, I calculate the Rouge scores (Lin, 2004) of the summaries generated by the three summarisation models. As there are multiple summaries for a single source text, a Rouge score of the model output is calculated using each human summary as a reference. The average of these values is used as the model's evaluation score. Table 4.6 shows the Rouge scores of the three

[6] https://huggingface.co/docs/transformers/v4.26.1/en/internal/generation_utils

	Rogue-1	Rogue-2	Rouge-L
CS19			
BART-PT	.450	.181	.273
BART-FT	.299	.079	.192
DistilBART	.303	.073	.191
IELTS33			
BART-PT	.395	.105	.191
BART-FT	.287	.057	.163
DistilBART	.267	.044	.153
SSG23			
BART-PT	.360	.099	.187
BART-FT	.274	.044	.148
DistilBART	.284	.045	.161

TABLE 4.6: Average Rouge values of the models.

models. The table shows the models' performance at initial top-k ablation where k is 0.

Figures 4.7, 4.9, and 4.10 and show macroscopic input ablation using free-decoding methods on CS19, IELTS33 and SSG23 datasets. The y-axis on those figures shows the Rouge score against the number of removed words, segments, and sentences at each generation step on the x-axis. The graphs show only the interpretation and aggregation method combination resulting in the highest correlation, along with baseline methods with the FC feature in terms of Table 4.5.

In the CS19 macroscopic ablation (Figure 4.7), the Attn method with max aggregation shows a similar decreasing curve shape with the fixation count (FC). Their ablation curve mostly overlaps from initial top-1 saliences to top-6 tokens, top-2 segments, and top-3 sentences. This overlap might explain their moderate and strong positive correlation at the token and sentence levels.

Overall, the macroscopic input ablation on CS19 shows a performance decrease. I observed a slight increase in the middle point of top-k saliencies, providing a nuanced view of the performance trend where the BART-PT Rouge-2 and Rouge-L at segment and sentence ablation. These increasing Rouge-score could be seen in details for each instance of source and participant summary in Figure 4.8. The figure shows a source text with two heatmap from BART-PT model with Attn saliency and FC feature from human. The index indicate saliency value of the model, the lower rank indicate higher saliency value with high intensity colour. The reference is the participant P09 summary of the source text. The table shows a run of macroscopic input ablation when removing top-k segments where $k = [0, 1, 2, 4, 5, 6, 7, 8]$ for the source text and summary pairs.

Figure 4.9 shows that removing the top-k tokens (Lead) increases the summarisation score within the ablation evaluation instead of decreasing the performance. These results might be related to the weak negative correlation to the human gaze (FC). Another interesting observation to note is that the FC

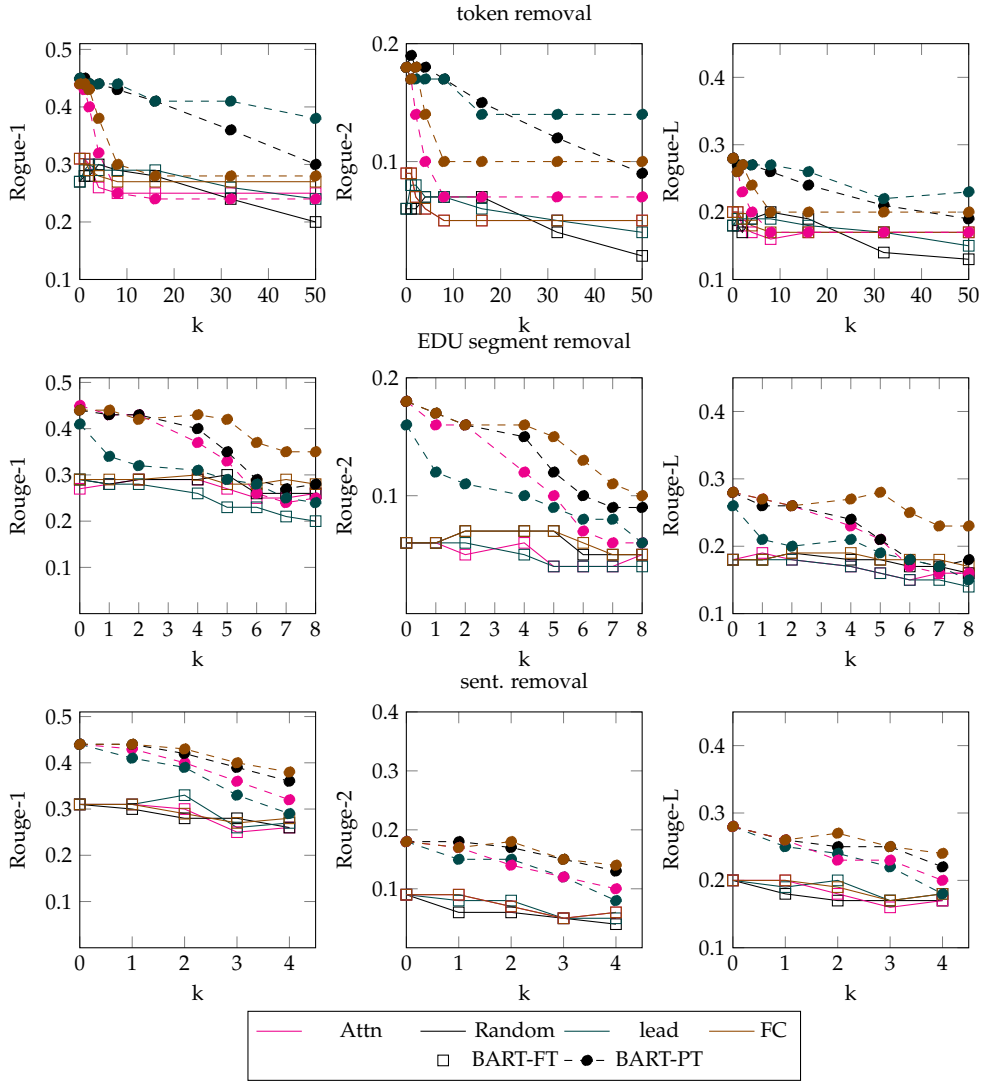


FIGURE 4.7: Free-decoding ablation analysis for CS19

ablation curve is lower or close to the BART-FT Attn and the BART-PT Attn. This result might indicate one of the positive answers to the RQ2 in this study for the summarisation task. However, the ablation at segment and sentence levels is almost flat except for the baseline methods, which instead are increasing.

Figure 4.10 presents similar trends, with the Lead method showing initial decreases at the three removal scenarios. However, the top-k token ablation at k=32 and k=50 shows an increased score for both BART-PT and BART-FT models. The FC ablation curve is also lower or close to the BART-FT Attn and the BART-PT Attn at token removal. Yet, there is an intriguing exception where the FC method removal is higher than random, raising questions about the uniform structure of the important information in the source texts or other factors that might be considered.

Overall, the answer to RQ2 using macroscopic ablation is positive at token level. There is a relation between the rank correlation (Table 4.5) and the

Model/Segment

6_[F am ilies hit with increase in cost of living British families have to cough up an extra Â£ 31 , 300 a year]

3_[as food and fuel prices soar at their fastest rate in 17 years . Pr ices in supermarkets have climbed at an alarming rate over the past year . Analy sts have warned]

5_[that prices will increase further still ,]

4_[making it hard for the Bank of England to cut interest rates]

1_[as it struggles to keep inflation and the economy under control . To make matters worse ,]

8_[escalating prices are racing ahead of salary increases , especially those of nurses and other healthcare professionals ,]

2_[who have suffered from the government _ s insistence]

7_[that those in the public sector have to receive below - inflation salary increases . In addition to fuel and food , electricity bills are also soaring . Five out of the six largest suppliers have increased their customers _ bills .]

High Low

Human/Segment

7_[F am ilies hit with increase in cost of living British families have to cough up an extra Â£ 31 , 300 a year]

6_[as food and fuel prices soar at their fastest rate in 17 years . Pr ices in supermarkets have climbed at an alarming rate over the past year . Analy sts have warned]

1_[that prices will increase further still ,]

4_[making it hard for the Bank of England to cut interest rates]

3_[as it struggles to keep inflation and the economy under control . To make matters worse ,]

5_[escalating prices are racing ahead of salary increases , especially those of nurses and other healthcare professionals ,]

8_[who have suffered from the government _ s insistence]

2_[that those in the public sector have to receive below - inflation salary increases . In addition to fuel and food , electricity bills are also soaring . Five out of the six largest suppliers have increased their customers _ bills .]

High Low

Reference: Inflation and rising interest rates are forcing companies to cut costs and raise prices in the UK. This is costing 30 thousand pounds extra per family and is not likely to decrease. The Banks and Government are unlikely to assist.

k	Attn Rouge-1	Attn Rouge-2	Attn Rouge-L	FC Rouge-1	FC Rouge-2	FC Rouge-L
0	0.137	0.028	0.110	0.137	0.028	0.110
1	0.189	0.028	0.108	0.240	0.027	0.133
2	0.189	0.028	0.108	0.187	0.027	0.187
4	0.263	0.054	0.105	0.187	0.027	0.187
5	0.164	0.028	0.137	0.187	0.027	0.187
6	0.179	0.031	0.090	0.110	0.028	0.110
7	0.103	0.000	0.103	0.119	0.031	0.119
8	0.103	0.000	0.103	0.119	0.031	0.119

FIGURE 4.8: BART-PT-Attn Macroscopic ablation on CS19-U2-P09 instance between Attn and FC in short segment.

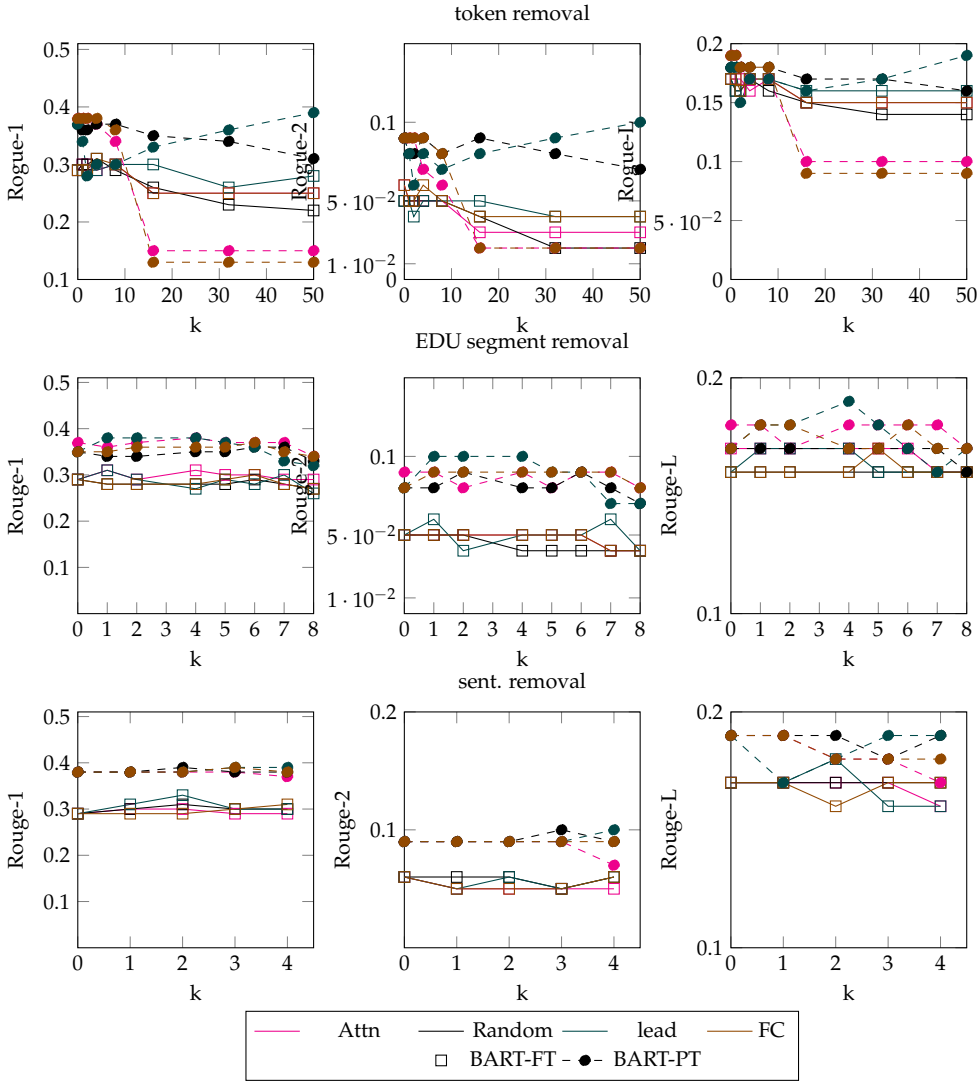


FIGURE 4.9: Free-decoding ablation analysis for IELTS33

performance. The evidences are based on the correlation sign and the ablation curve's increase or decrease (Figures 4.7, 4.9, 4.10). However, I do not see a positive relation between rank correlation and the performance at segment or sentence level.

4.7 Microscopic Analysis

This section analyses the relationship between the model saliency and fixation counts on words at each step of generating a word in summaries. Aligning these two values is the first step in investigating the relationship between a sequential saliency matrix and a temporal-segment eye-gaze matrix. The temporal segment needs to be created from eye-gaze data. However, a duration corresponding to each word in a summary must be defined. To define a temporal segment for a word, I propose two methods: fixed-number segmentation (Fix) and keystroke-based segmentation (Key).

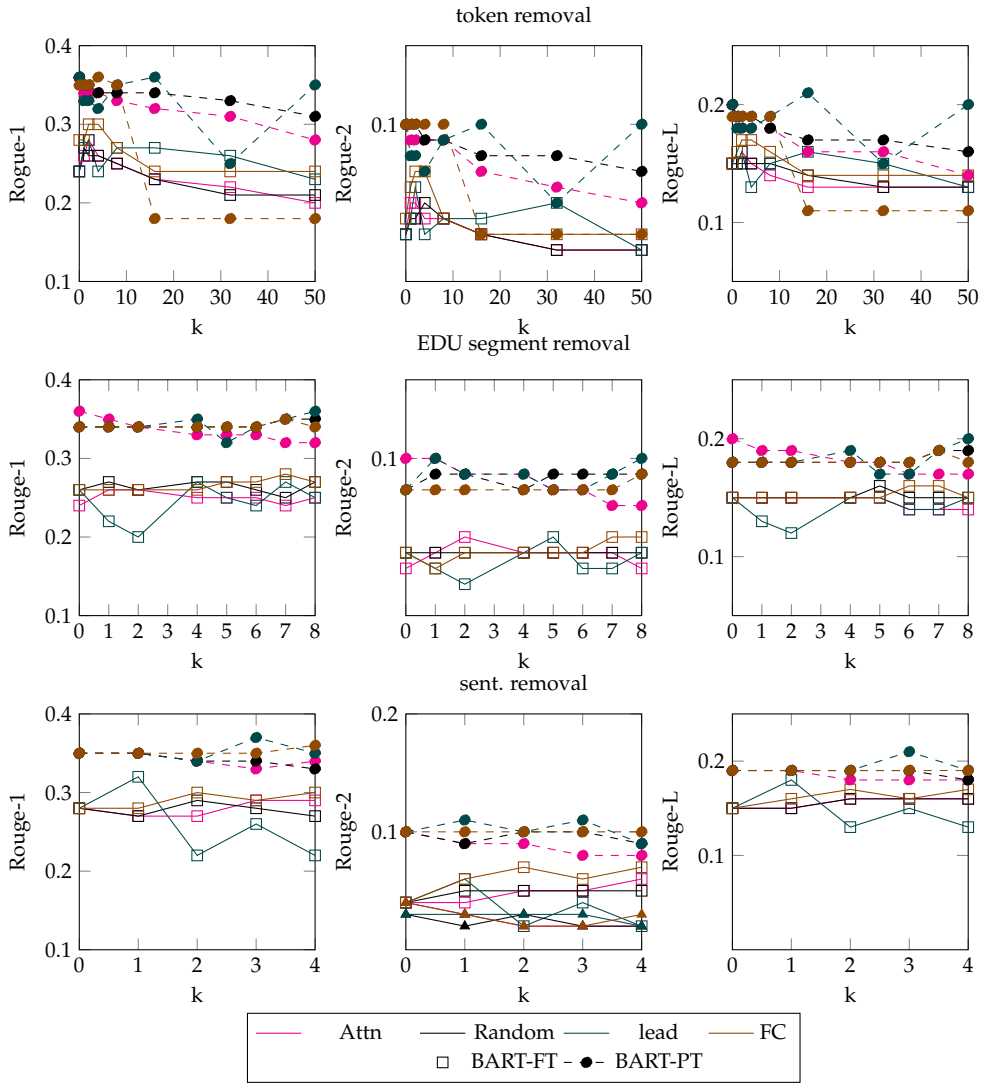


FIGURE 4.10: Free-decoding ablation analysis for SSG23

Fix-segment The fixed-number segmentation follows the piecewise approximation aggregation (PAA) (Keogh et al., 2001), which segments a time series data by dividing them into temporally equal-sized segments; the value of each segment is calculated by averaging values in the segment. In our case, I divide a sequence of fixation counts for the entire summarisation process into m segments of equal duration and sum up the fixation counts in each segment. In this method, however, each segment is not guaranteed to align with the generated word.

Keystroke-segment To realize a more precise alignment between a temporal segment and a generated word, I introduce keystroke-based segmentation. This study proposed a temporal segment for a generated word as a duration between the first and last key input time points of the word. Participants might edit the word until they finalise it; I include editing time to the temporal segment.

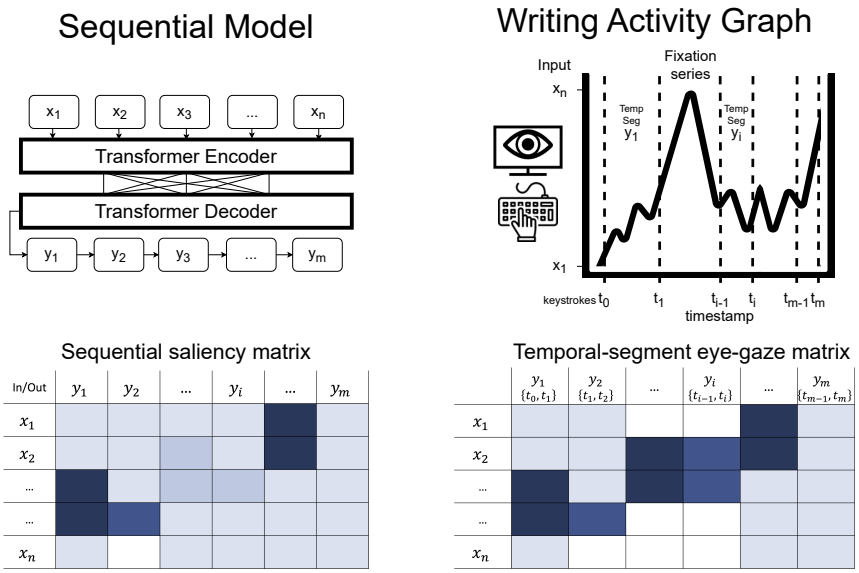


FIGURE 4.11: Sequential saliency matrix (lower-left) and temporal-segment eye-gaze matrix (lower-right). Dense colour represents high saliency and frequent fixation counts, respectively.

The fixation counts are summed up in each segment. The right of Figure 4.11 illustrates the keystroke-based segmentation and the resultant fixation count matrix.

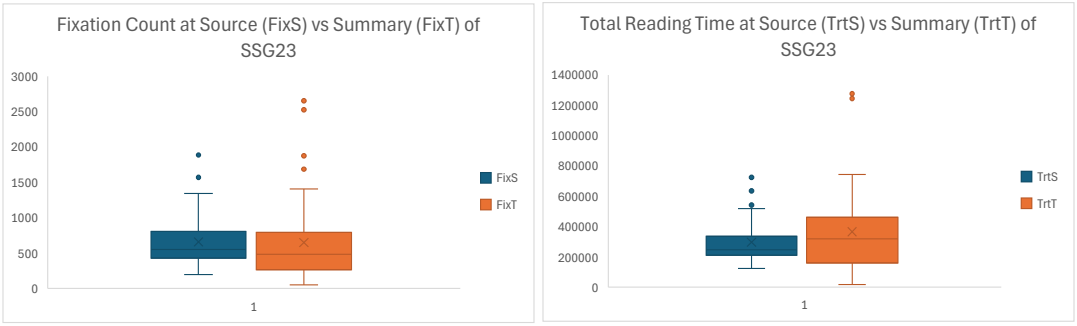
4.7.1 Discussion for RQ1

I propose calculating the rank correlation between model saliency and human fixation counts at each step of generating a word. Given a source text, a saliency vector is obtained (a column vector in the left-bottom matrix in Figure 4.11) at each step of generating each human summary by force decoding. For a human summary of m word tokens, there are m saliency vectors to construct a saliency matrix shown in the left bottom of Figure 4.11. Unlike the macroscopic analysis, it is not possible to aggregate the saliency matrices across the participants because each participant's summary length (m) is different. Therefore, I use Spearman's rank correlation between each pair of individual saliency vectors (a column in the left-bottom of Figure 4.11) and the corresponding fixation vector (a column in the right-bottom). They are averaged over m words in a summary and averaged over the participants and source texts. Similar to the macroscopic analysis, the instance correlation value is transformed into the Fisher's z score and converted back to ρ value (Myers and Sirois, 2006) to calculate the averaged rank correlation.

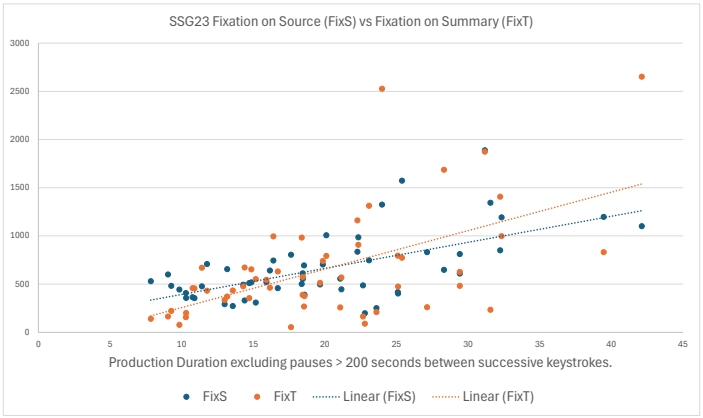
Table 4.7 shows the average rank correlation between model saliency and fixation counts. Similar to the macroscopic analysis, Attn shows a stable and slightly stronger correlation than other interpretation methods at the sentence level. However, compared with the result of the macroscopic analysis

Dataset\Seg.	Fix			Key		
Model+Interp.	token	seg.	sent.	token	seg.	sent.
CS19						
Random	-.002	.007	.044	-.002	-.011	.064
BART-PT Grad	.038	.163	.155	.067	.254	.427
BART-PT IG	.016	.110	.042	.057	.202	.296
BART-PT Occ	-.002	-.066	.011	.037	-.035	-.001
BART-PT Attn	.114	.185	.201	.180	.297	.467
BART-FT Grad	.075	.178	.282	.111	.303	.525
BART-FT IG	.054	.128	.280	.074	.255	.506
BART-FT Occ	.016	.032	.060	.030	.045	.094
BART-FT Attn	.115	.179	.297	.184	.334	.553
IELTS33						
Random	-.004	.003	-.007	.000	-.003	.018
BART-PT Grad	-.009	.027	.078	.006	.058	.206
BART-PT IG	-.002	.014	.058	.009	.036	.190
BART-PT Occ	-.003	-.044	-.006	.010	-.013	.018
BART-PT Attn	.032	.090	-.050	.052	.092	.054
BART-FT Grad	-.008	.036	.074	.006	.058	.176
BART-FT IG	-.005	.035	.045	.008	.058	.146
BART-FT Occ	-.005	-.006	-.031	.014	.012	.029
BART-FT Attn	.024	.060	-.012	.041	.069	.063
SSG23						
Random	-.002	.002	-.001	-.001	.005	-.004
BART-PT Grad	.042	.248	.289	.013	.070	.092
BART-PT IG	.022	.228	.281	.013	.059	.071
BART-PT Occ	-.001	-.004	.012	.006	.011	.007
BART-PT Attn	.027	-.042	-.035	.036	.108	.099
BART-FT Grad	.031	.224	.229	.016	.077	.083
BART-FT IG	.015	.211	.247	.017	.071	.098
BART-FT Occ	.001	.064	.013	.006	.014	.027
BART-FT Attn	.009	-.088	-.100	.028	.096	.072

TABLE 4.7: Average rank correlation at each word generation.



(A) Fixation count and total reading time distribution from source and summary in SSG23 dataset.



(B) Production duration (x-axis) and Fixation count on Source vs Summary (Y-axis).

FIGURE 4.12: Fixation Count on Source vs Summary on SSG23 dataset.

(Table 4.5), the correlation coefficients are significantly low, particularly at the token level.

In most cases, model have a stronger correlation to the keystroke-based segmentation (Key) compared to the fixed-number segmentation (Fix). However, there is an exception on SSG23 where the models are more correlated to the Fix-segment than the Key-segment. This might be a special case because there are more fixation counts on summary than source in SSG23 (Figure 4.12a) unlike the other two datasets (Figures C.1a and C.2a). The scatterplot figures 4.12 show a relation between fixation count on source (FixS) and production duration and also fixation count on summary (FixT) to production duration. Production duration is the total duration of successive keystrokes. The Production duration is the total duration of successive keystrokes. It is related to keystroke segmentation. At a close time, a fixation can only happened either at source or summary but not both at the same time. The plot imply that when typing a summary, the participants fixated on the summary than the source text during their typing, which causes the keystroke duration to capture only a small portion of the fixation on the source. In contrast, binning captures the fixation on the source outside the keystroke duration. The scatter plot on appendix C shows that production duration is more related to source than summary on CS19 and IELTS33.

These results show that the models' generation process aligns better with

keystrokes than the fixed-size duration, which assumes humans write linearly in time without backtracking. However, when fixation on summary is more than fixation on source while writing, fixed-number segmentation can capture fixation during writing better than keystroke-based segmentation.

The answer to RQ1 is that the experiment results show weak correlations between word saliency from the interpretation method and fixation counts on the source text sentences under some conditions. Similar to macroscopic analysis, the attention-based interpretation method (Attn) is promising.

4.7.2 Discussion for RQ2

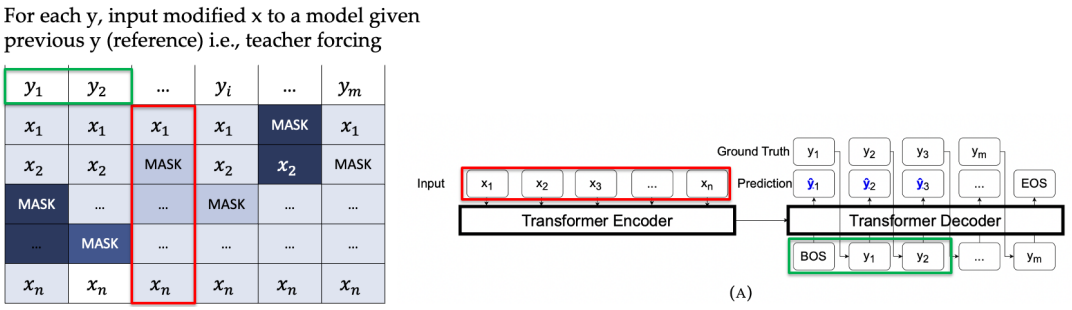


FIGURE 4.13: Microscopic features ablation using force-decoding

At the microscopic level, I also proposed to apply the input ablation method, which was used to evaluate model saliency conformity relation to its performance in NLP tasks (Jacovi and Goldberg, 2020). The method has been applied to evaluate saliency in NLP models (DeYoung et al., 2020; Xu and Durrett, 2021).

Following Xu and Durrett (2021), the ablation evaluation replaces the top k salient words with a special mask token at each step of generating a word. From figure 4.13, it shows there is an ablated text for each step prediction. For each step, the input to the models is a modified text (enclosed by a red line in Figure 4.13) and a previous ground-truth output (enclosed by a green line in Figure 4.13). The output is a log probability distribution at each step. To select an output token, Xu and Durrett (2021) use arg max and observe the loss value against the ground truth for prediction at each step. However, greedy decoding still makes the model output nonsense text despite the force-decoding scenario. This shows that the greedy decoding in training is suboptimal and does not reflect the model prediction well. Therefore, I apply beam search decoding for the model output. This beam search decoding makes the output an upper bound of the model inference and more similar to the human summary. However, it does not necessarily mean the model output is the same as the output token to references. Also, I use the Rouge score of the generated summaries instead of the model loss because Rouge scores indicate the summary quality more directly. In the evaluation, I use the Rouge scores as in Section 4.6.2, i.e., using human summaries as the reference summary.

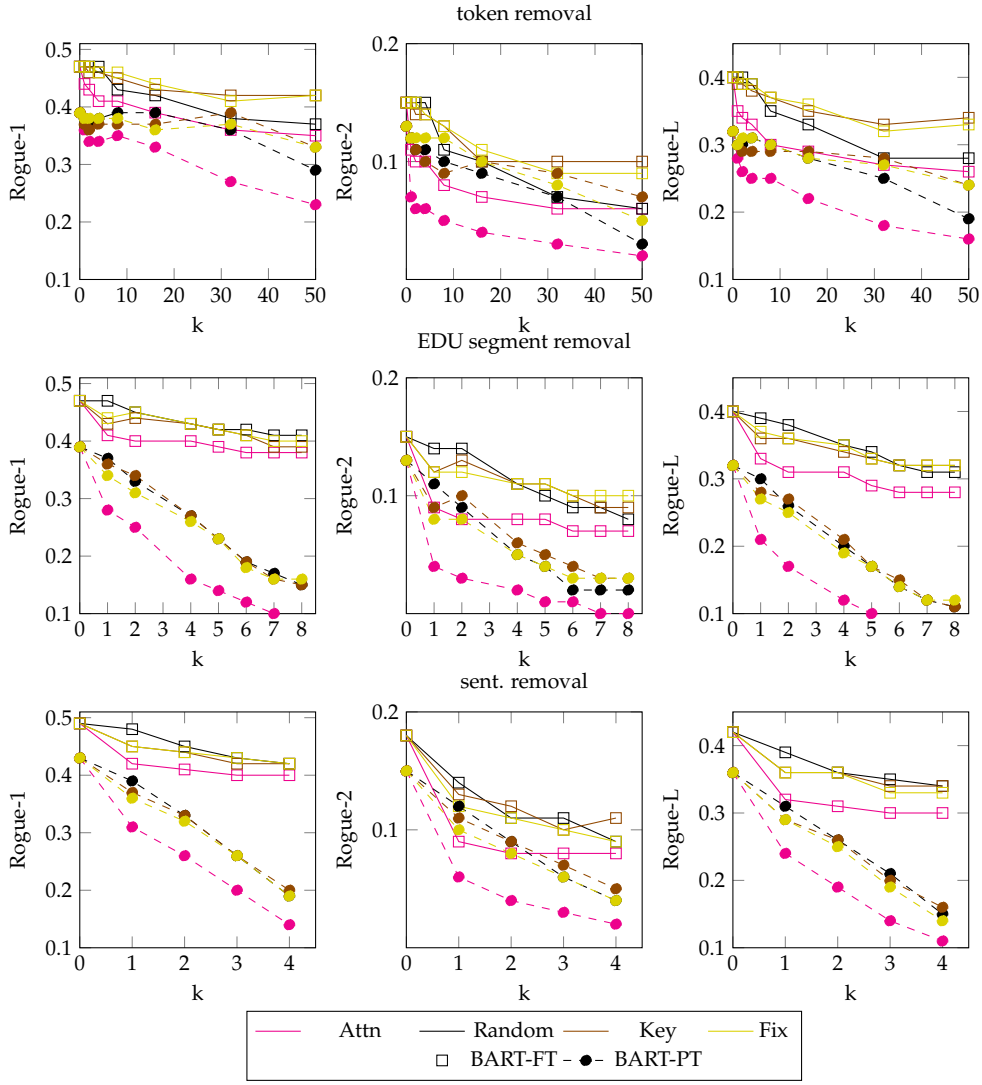


FIGURE 4.14: Ablation analysis for CS19

I also conduct the word ablation based on the fixation count and compare the change of Rouge scores between model saliency and human fixation counts. In addition, I conduct EDU segment and sentence ablation, where top k salient segments or sentences are simply removed from the input.

Figure 4.14, 4.15, and 4.16 show the Rouge score (y-axis) against the number of removed words, segments and sentences at each generation step (x-axis). To avoid the diagrams becoming complicated, I only show the results of the model saliency-based ablation (Attn), a baseline (Random), and the fixation-based ablation (Key and Fix). Comparing three datasets, the ablation impacts the Rouge scores more mildly in IETLS33 and SSG23 than CS19. Particularly, the difference is significant in the segment or sentence ablation. This happens due to longer source texts in the IELTS33 dataset.

BART-FT shows consistently higher Rouge scores than BART-PT, which is the opposite result of the summary generation with free-decoding (Table 4.6). However, this result conforms with the higher correlation of BART-FT than BART-PT in the macroscopic analysis.

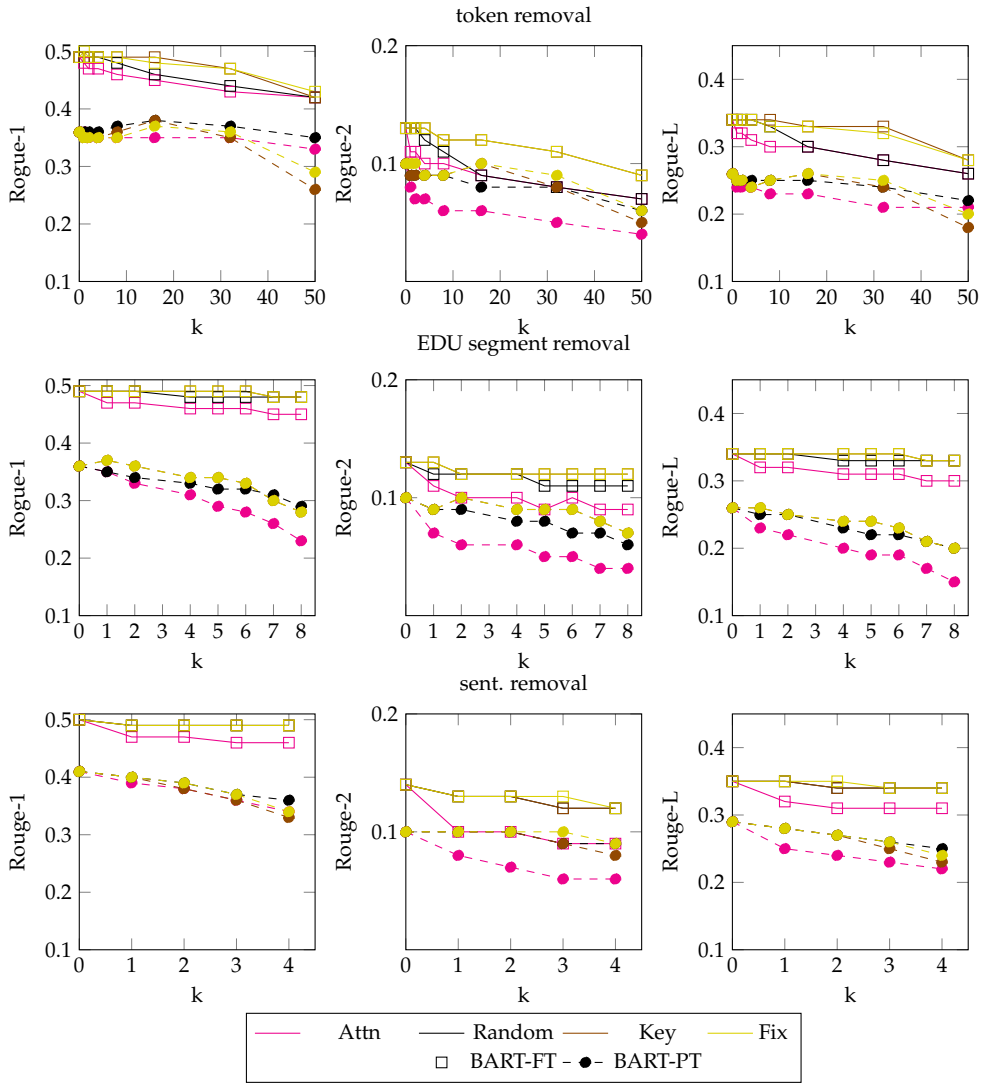


FIGURE 4.15: Ablation analysis for IELTS33

There is no much difference between the temporal segmentation methods: fixed-number segmentation (Fix) and keystroke-based segmentation (Key) in these graphs. Similar to macroscopic ablation, the ablation curve increasing or decreasing is more important than the plot absolute value. Concretely, the Attn, Key blation curve which show a similar decrease at k removal shows the models' saliency might select similar part of texts in most of the instances.

From Figure 4.16, the fixed-size human gaze (Fix) at token-level reduce the model Rogue scores similarly to the Attn method and the Key gaze. The reason might be the fact that the interpretation method more correlates to the Fix feature than to the Key feature in the SSG23 dataset (Table 4.7).

The force-decoding ablation has the same impact on different datasets and interpretation methods. Additionally, it does not show any relationship between microscopic correlation, representing conformity. One reason is that force-decoding methods do not reflect the model prediction phase and always see the correct output of the previous step. Another possible reason

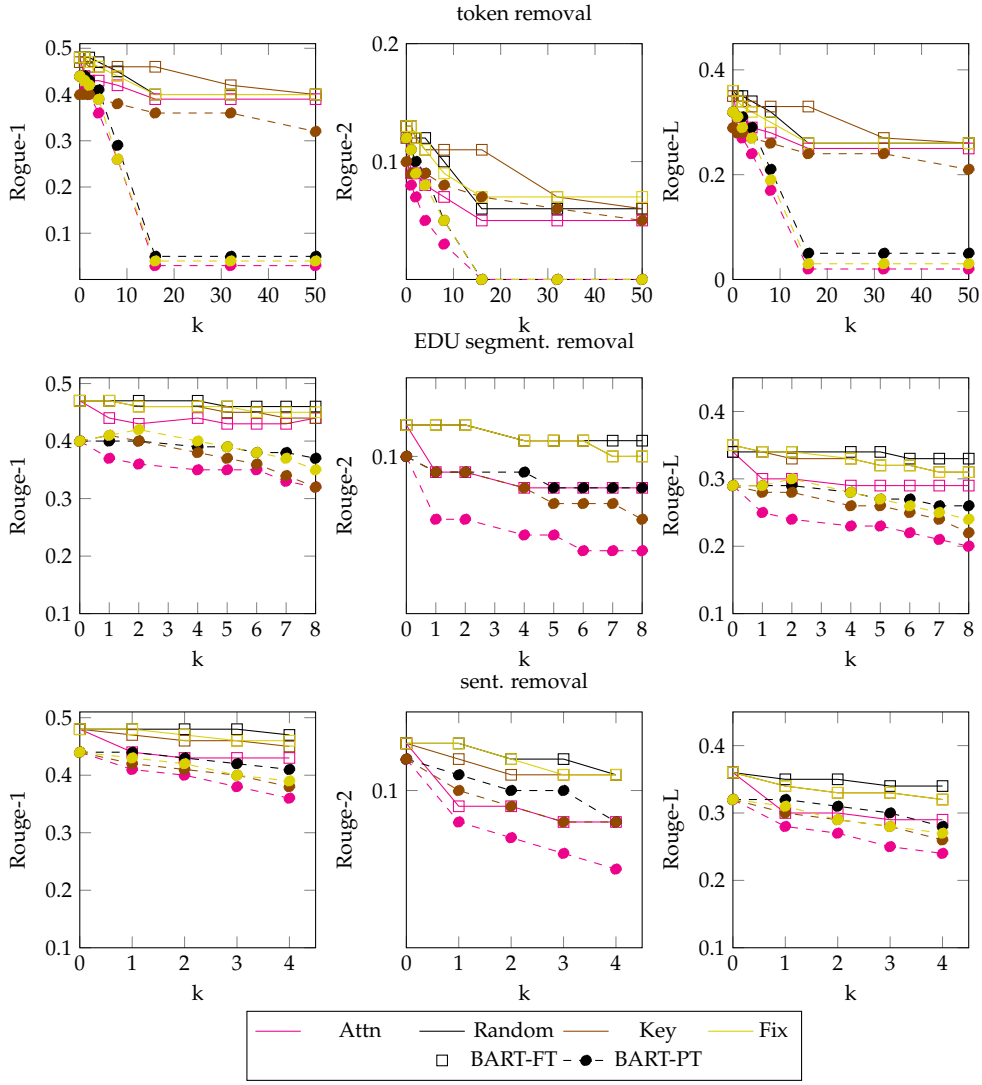


FIGURE 4.16: Force-decoding ablation analysis for SSG23

is that force-decoding methods make the models heavily depend on the decoder input when generating a text (Voita, Sennrich, and Titov, 2021).

The ablation method with free-decoding offers a more realistic scenario than force-decoding results. This shown by the performance consistency when evaluating the model performance (Table 4.6) where at k=0 in force-decoding ablation. The table shows that BART-PT performance is higher than BART-FT in free-decoding. However, the model evaluation when using force-decoding ablation shows BART-FT is higher than BART-PT. The performance measure in force-decoding might be related when we see training loss is lower than test loss. This is called exposure bias (Wiseman and Rush, 2016).

To summarize, the answer to RQ2 in this microscopic analysis is negative.

4.8 Chapter Summary

In this chapter, I proposed a novel framework for analyzing summarisation models by comparing them to the human summarisation process with eye movement as a proxy. The framework comprises macroscopic and microscopic analyses of model saliency and human gaze data. In macroscopic analysis, I compared the model saliency and eye gaze at the input level. In microscopic analysis, I compared the model saliency and eye gaze at each output token. To align model saliency and eye-gaze information at every token generation, I introduced temporal-based segmentation for the time series of fixations.

This study answered two research questions using the framework. RQ1: *Does the input word saliency from interpretation methods in summarisation models conform with human eye-gaze features during summarisation?* RQ2: *How does the model saliency conformity impact model prediction?*.

According to the correlations between model saliency scores and human fixation counts in the macroscopic and microscopic analyses, our answer to RQ1 is partially yes, particularly for the attention-based saliency scores. The macroscopic ablation analysis showed that removing important words according to the human gaze can reduce performance as much as sometimes higher than model attention. The microscopic ablation analysis showed that removing important words according to the human gaze did not degrade performance as much as the removal by the model saliency. The microscopic ablation does not align with humans because the forced decoding methods might not reflect the prediction scenario and the possible model bias towards the decoder input, which is the previous ground-truth steps. Thus, the answer to RQ2 is also partially yes. The human gaze might affect the models in typical prediction. However, forced decoding methods make the model depend on previous ground truths. The force-decoding ablation showed the models' upper-bound performance, but it does not reflect well on humans.

Chapter 5

Conclusions

Deep learning's interpretation methods provide the information the machine considers essential. At the same time, the human eye gaze has been believed to be a proxy for the human cognitive process. In this thesis, I presented a comprehensive series of studies from interpreting the reading tasks to the writing process of deep learning in natural language processing tasks using eye gaze information as a cognitive proxy. It is motivated by the need for understanding of the complex models' decision making process from its deep layer. This study could also be used to improve model robustness and efficiency of training data by taking inspiration from human cognitive process.

This thesis includes two studies :

Ch3. Interpreting models in comparison to eye gaze in reading tasks.

Ch4. Interpreting models in comparison to eye gaze in summarisation.

These chapters are connected studies through the increasing complexities of both factors: from reading to writing tasks in humans and classification to generation tasks in machines.

Specifically, this thesis aims to answer the following research questions.

RQ1: Does the input word saliency from interpretation methods conform with human eye gaze features?

RQ2: How does the model saliency conformity impact model prediction?

RQ1 concerns whether the machine looks at the same input elements as the human to solve the task. RQ1 leads to RQ2, which concerns whether the machine which behaves like humans performs the task better.

5.1 Interpreting Models on Reading tasks.

In the first study, I investigated a broad overview of the relation between different interpretation methods and human eye-movement behaviour across different tasks and architectures. I analyse three types of natural language processing (NLP) tasks: sentiment analysis, relation classification, and question answering, and four interpretation methods: simple gradient, integrated gradient, input-perturbation, and attention, with three architectures: LSTM, CNN, and Transformer. I utilise two publicly available corpora annotated with eye gaze information: the ZuCo and MQA-RC datasets.

To discuss RQ1, I introduced the saliency distance (SD), defined as the KL-divergence between the saliency distributions over input words by an interpretation method and an eye gaze feature; a small SD score indicates good conformity between human visual attention and focal words by machines. The analysis results with SDs provided a deeper understanding of the relationship between saliency distance and various factors and offered practical insights. The degree of conformity, which varied depending on the combinations of the tasks, interpretation methods, and architectures, and each question instance, has direct implications for the design and implementation of machine interpretation methods in real-world scenarios. Concretely, transformer models generally showed small SD scores regardless of the tasks; the gradient-based interpretation methods showed small SD scores for the tasks with relatively short input sentences (SA, RC and QA-Span). Then, I analysed the the precise differences in the saliency distributions between humans and machines which support our claims. This findings and analysis account for the chapter's (Chapter 3) first contribution.

Concerning RQ2, I investigated the difference between average SD scores over correct and wrong test instances. A high similarity (a low SD score) with the human score only sometimes led to a correct answer. I proposed the SD-performance curve (SDPC) for detailed analysis, representing cumulative model performance against the SD scores. The SDPC not only uncovers hidden phenomena that are often overlooked by macroscopic metrics, such as average SD scores and rank correlations, but also provides a practical tool for understanding and improving model performance in real-world applications. For instance, SDPC told us that Transformer showed an initially decreasing trend of SDPCs for the QA-MC task, even though their average SDs over inside and outside answer spans were not significantly different. The SDPC proposal constitutes the chapter's (Chapter 3) second contribution.

5.2 Interpreting Models in summarisation

In the second studies, I proposed a novel framework for analyzing summarisation models by comparing them to the human summarisation process with eye movement as a proxy. The framework comprises macroscopic and microscopic analyses of model saliency and human gaze data. In macroscopic analysis, I compared the model saliency and eye gaze at the input level. I compared the model saliency and eye gaze at each output token in microscopic analysis. To align model saliency and eye gaze information at every token generation, I introduced keystroke-based segmentation for the time series of fixations.

The results show correlations between model saliency scores and human fixation counts in the macroscopic and microscopic analyses. The answer to RQ1 is partially yes, particularly for the attention-based saliency scores.

The microscopic ablation analysis indicated that removing important words based on human gaze did not degrade performance as much as removal based on model saliency. This discrepancy is likely due to forced decoding methods not accurately reflecting the prediction scenario and potential

model bias towards the decoder input, which relies on previous ground truth steps. Thus, the answer to RQ2 is partially affirmative. While the human gaze may influence models in typical prediction scenarios, forced decoding methods cause the model to depend heavily on previous ground truths. Consequently, although forced decoding ablation demonstrates the models' upper-bound performance, it does not align well with human behaviour.

5.3 Cognitively plausible Models in Large Language Models (LLMs) Era

In recent years, we have seen the advancement of pre-trained large language models. Although no exact definition exists, these models typically have parameters larger than billions of parameters (Chowdhery et al., 2023; Touvron et al., 2023). These models exhibit remarkable capabilities such as multi-task learning (Radford et al., 2019) and few-shot learning (Brown et al., 2020). These models are typically trained by predicting the next-word tokens from vast amounts of text data. This advancement has sparked a growing interest in investigating the cognitive behaviour of LLMs, which achieve high performance on many benchmarks (Chiang et al., 2024; Hendrycks et al., 2020; Wang et al., 2019) but fail on reasoning and logical benchmarks (Dziri et al., 2023).

Despite their impressive performance, LLMs often function as opaque systems, lacking transparency (Bommasani et al., 2023). Recent study (Oh and Schuler, 2023) also shows that larger language models diverge when estimating human-like surprisal. Subsequent study (Oh, Yue, and Schuler, 2024) attributes this behaviour to inverse correlations of models' predictive capabilities to the frequency of rare words from Internet data. The models learn to predict rare and complex words better than the average human ever encounters. The models' parameters and training data size make LLMs a poor choice for cognitively inspired language models. Another study (Kuribayashi et al., 2022) showed that limiting the context size of language models increased the model fit to human reading behaviour. There are many benefits of integrating cognitive behaviour into NLP models.

Considering cognitive studies on language models, this thesis contributes to the investigation of natural language processing and cognitive behaviour. I proposed using eye gaze beyond reading tasks and predictive modelling by using interpretation methods. There are several possible directions for future work.

- **Interpretation methods**: Current experiments compared a limited representation of interpretation methods; other methods should be also considered. For instance, a recent study by Eberle et al. (2022) reported the attention flow (Abnar and Zuidema, 2020) method shows higher

correlations to human eye movements in sentiment analysis and relation extraction tasks. Although they do not investigate on any writing task. Reverse engineering multi-head attention in transformers architecture demonstrates emergent behaviour such as in-context learning (Elhage et al., 2021; Olsson et al., 2022). It is interesting to compare these methods with human cognitive process.

- **Eye-gaze datasets & models**: Collecting eye gaze data is indeed time consuming. However, there have been attempts to model reading behaviour. Early study (Sood et al., 2020a) utilised EZ-reader (Reichle et al., 1998) [1] to augment their gaze-guided paraphrasing models. Recent studies (Deng et al., 2023a; Khurana et al., 2023) proposed to predict scanpaths from eye-tracking data. They showed (Deng et al., 2023b) the feasibility of integrating the pre-training objectives to BERT models to generate synthetic scanpaths data. These methods can be integrated with current LLMs models to improve human-like behaviour and robustness.

- **Second Language learning**: Recent study (Oba et al., 2023) explored a scenario where non-English L1 pre-trained language models in acquiring English as a second language in grammatical knowledge. Their proposed methods could considered to expand to the summarisation task in second language learning of human and language models.

- **Application to other text generation tasks**: The proposed methods in this study can be easily extended to other tasks, such as paraphrasing or translation. Primarily, there are collections of publicly available translation data accompanied by eye-tracking data (Carl, 2012a). This dataset can be utilised to expand the model interpretation studies to the translation process.

[1] http://www.erikdreichle.com/downloads.html

Ethics and Data Privacy

In this study, I utilized publicly available eye-tracking datasets like ZuCo and Movie-QA. These datasets have been carefully designed to address ethical concerns. They are primarily used for machine learning and natural language processing training, as well as for studying the human reading process. This thesis objectives align with these uses. The data collection experiments for IELTS33 and SSG23 were reviewed and approved by the Ethical Review Committee of the author's university in advance. Prior to the experiment, the purpose and method of data collection, and usage of the collected data were explained to the participants, and their consent to participate in the experiment was obtained. The collected eye-tracking in this study does not hold any privacy information. I declare there are no competing interests regarding the funding of the data collection and this study.

Appendix A

Interpreting Models in Reading Tasks

A.1 Fixation Count SD-Performance Curve with Fixation Count

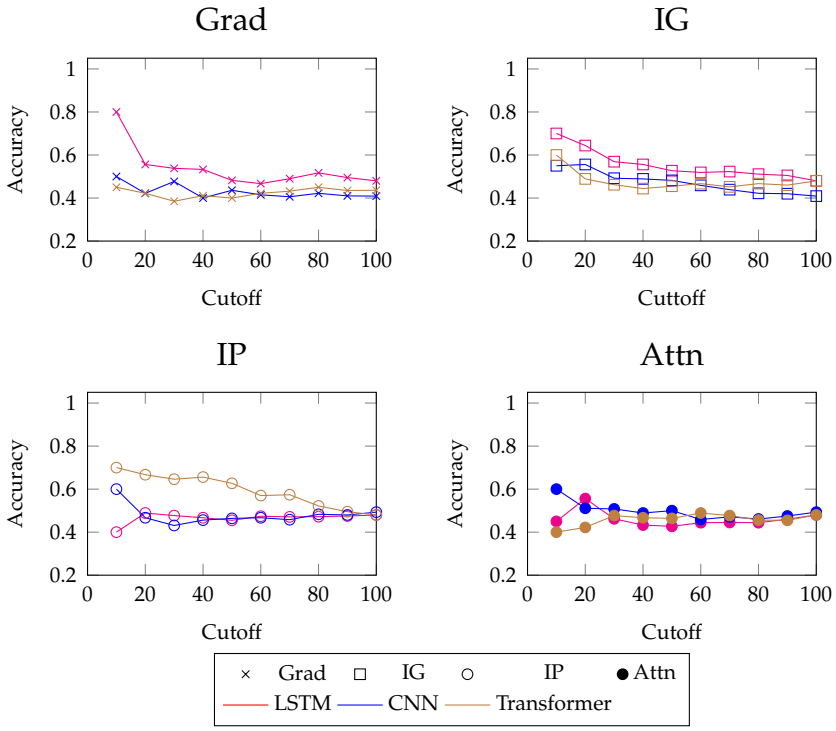


FIGURE A.1: SD-Performace curves for sentiment analysis (SA) with Fixation Count feature

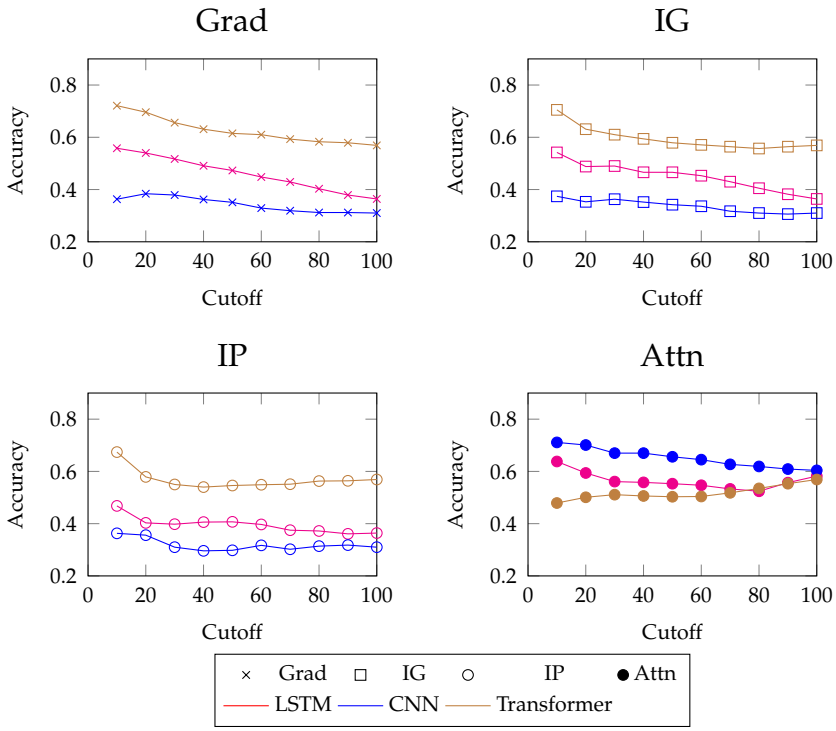


FIGURE A.2: SD-Performace curves for relation classification (RC) with Fixation Count feature

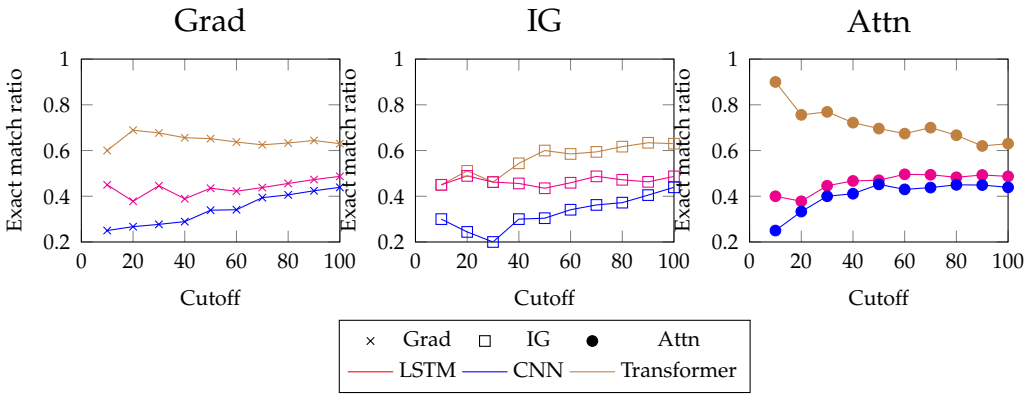


FIGURE A.3: SD-Performance curves for span-based QA with Fixation Count feature

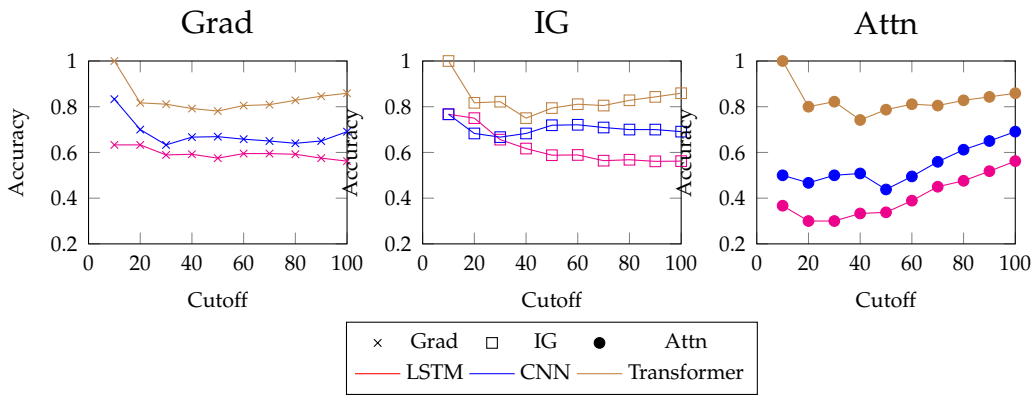


FIGURE A.4: SD-Performance curves for multiple-choice QA with Fixation Count feature

Appendix B

Eye-tracking Experiment Instruction

B.1 Setup

First, the Tobii Pro X3–120 Eye Tracker is used in a setup where it is attached to a screen and placed below the monitor. Open the Eye Tracker manager application (Figure B.1). You will see the list of available connected Eye Trackers. The list of connected Eye Trackers is displayed in the manager app, the currently-selected Eye Tracker is highlighted.

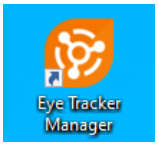
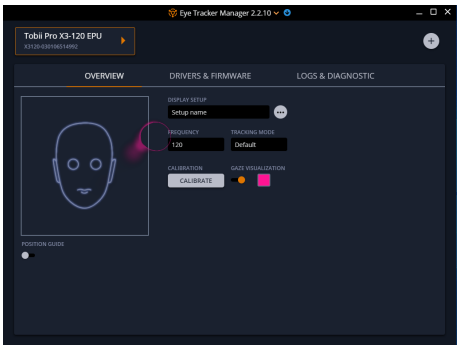


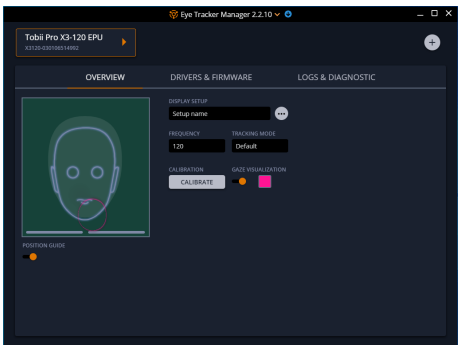
FIGURE B.1: Eye Tracker Manager application

B.1.1 Head Position

You can adjust your position comfortably by enabling the **position guide** which can be seen in figure B.2b.



(A) Position guide toggle disabled



(B) Position guide toggle enabled

B.1.2 Calibrate Device

Click on “Calibrate” on the Eye-tracker manager. This will start the calibration process so that the eye tracker knows where the participant is looking on the screen. The participant must focus on a series of dots on the screen. If all goes well, this should result in a 9-point square grid.

B.2 Translog-II

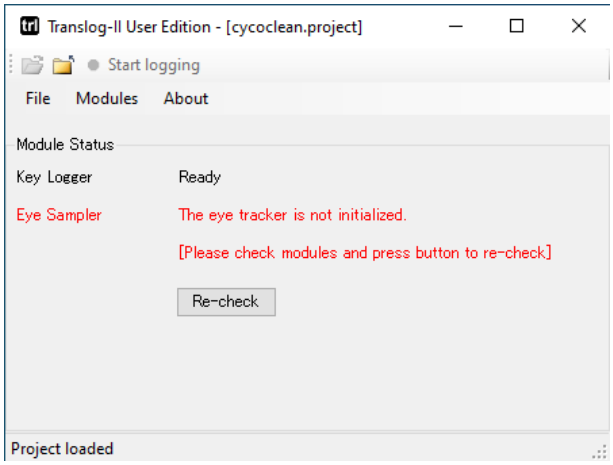


FIGURE B.3: Uninitialized eye-tracking device

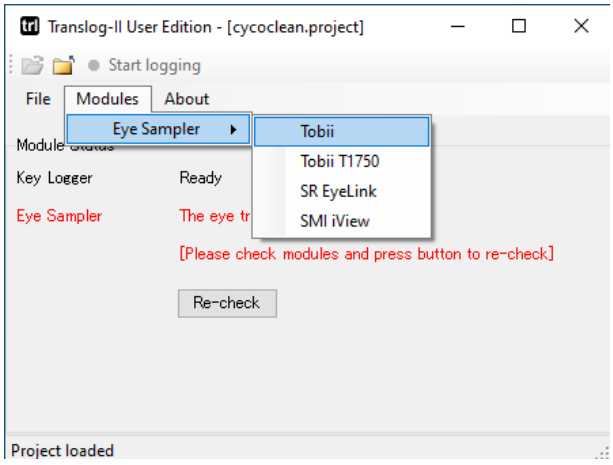


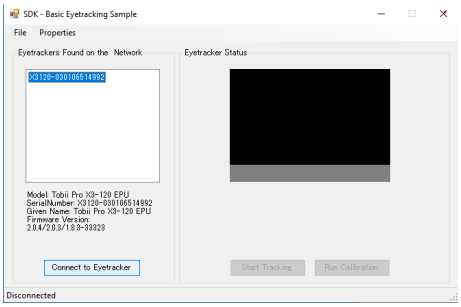
FIGURE B.4: choose device

B.2.1 Open Project

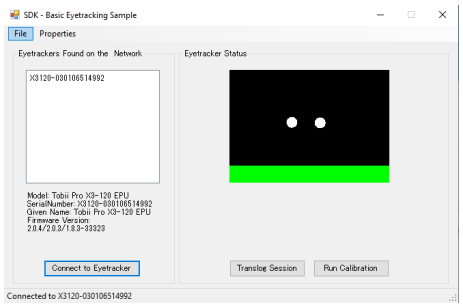
Open **Translog II User** application. Click **File** to open a project file and choose **Cycoclean** experiment. We will run an experiment involving eye-tracking, the window will appears after opening a project may look like Figure B.3.

B.2.2 Connnect Translog II with Eye-tracking device

"The eye tracker is not initialised" in the beginning, please select the Eye sampler with which you will be working. Navigate to modules eye sampler and select the eye tracker **Tobii** from the list (Figure B.4). The computer will automatically detected the eyetracker. Click on its code and, then, on "**Connect to Eyetracker**" (Figure B.5a). After connecting to eyetracker, you also need to



(A) connect to eyetracker, Click on its code and, then, on "**Connect to Eyetracker**" (Figure B.5a)



(B) connected to eyetracker, adjust head/eye position

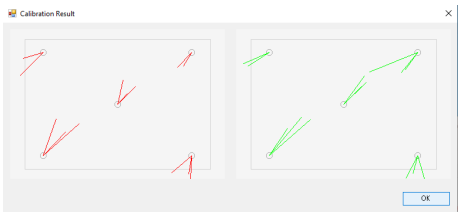
adjust your head/seat position from Translog-II application. Make sure your eyes are in the middle of the black screen and the rectangle below is green (unless you blink) and click "**Run Calibration**".

B.2.3 Calibrate Translog II with Eye-tracking device

FIGURE B.6: Yellow dot calibration

The calibration task consists in following a target with your eyes, usually focusing on its centre. The target are look like this **yellow dot** (Figure B.6). If the green and red lines cross all of the circles and short, the calibration was successful. To proceed, select "ok" in this instance. If "not enough data to create calibration" appears or the lines are very long and dispersed, the calibration was unsuccessful. repeat the procedure from step 2.2 .

(A) Good calibration result where the green and red lines cross all of the circles



(B) Poor calibration result where lines are very long and scattered. **Repeat the process.**

B.3 Running Experiment

Please follow the guide from the researcher to do eye-tracking calibration. When asked for participant/subject name at the start of input program, please write your experiment ID number, example TT01 for ID 01.

After the calibration phase, the eye sampler will be visible as "ready". Click on "Start Logging" to proceed. **Please finish this experiment within 50 minutes or less**, excluding reading this instruction and calibrating the eye-tracker.

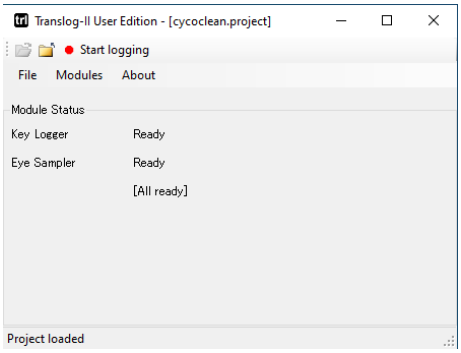


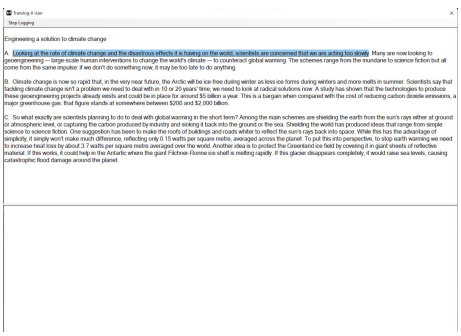
FIGURE B.8: After calibration phase, start logging to start the experiment

B.3.1 Reading

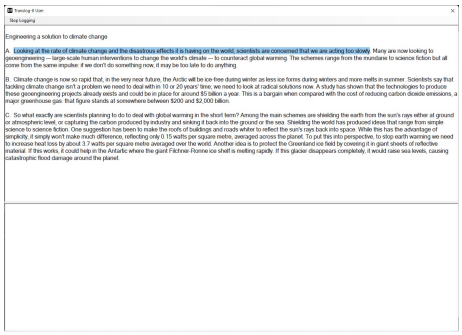
Please read the text thoroughly and understand the content of the essay. **The tool has an underline feature. Please use it to to mark the important sentence. To highlight a phrase or a sentence please drag and select the words you want to underline and press mouse right click button.** After you finish reading and marking all the important sentence you can find, please start writing the summary of the text.

B.3.2 Summarisation

Please create a summary of the text's most important points. **This summary should not exceed 5 lines of the summary windows at the bottom of the screen. The summary is limited to around 80 words**. In the summary, do not add additional information or personal opinion to the text, only use information from the text. The summary is intended for general audience where



(A) add highlight by pressing mouse button on selected text



(B) remove highlight by pressing mouse button on the underlined texts

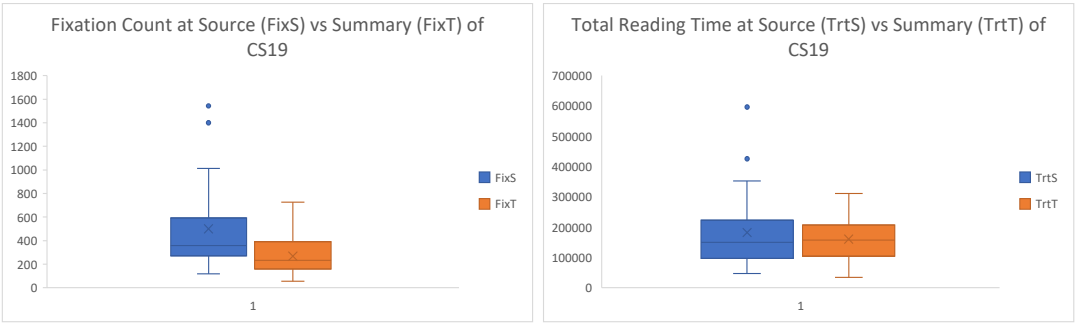
you will describe information what was contained in text. As for the wording of your summary, you can use your own formulations if you feel they are more appropriate or stick closely to the source text.

B.3.3 Finish

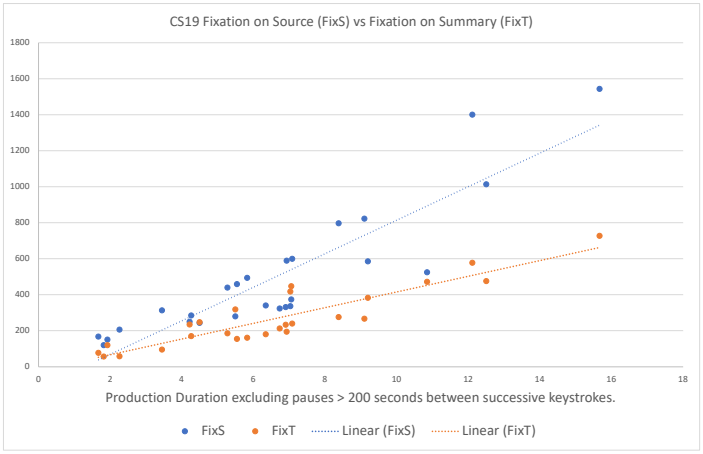
After you finish writing the summary, please select **stop logging** and save the file as {TT<ID>}-{mmddyy}-U1.xml file in the save menu.

Appendix C

Eye-gaze Summarization Data Analysis

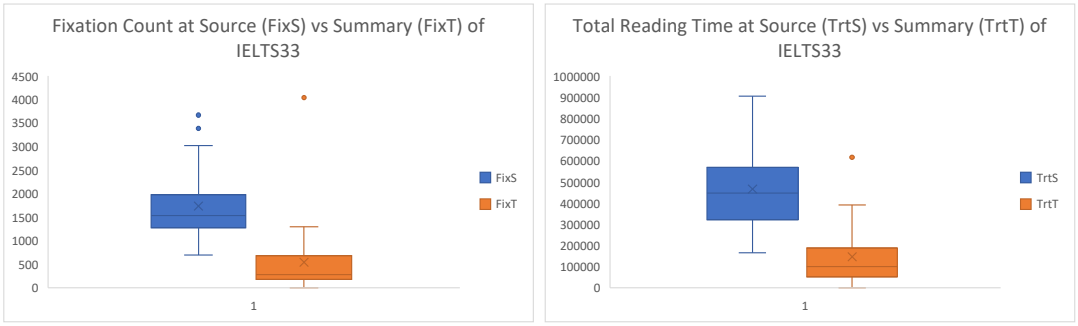


(A) Fixation count and total reading time distribution from source and summary in CS19 dataset.

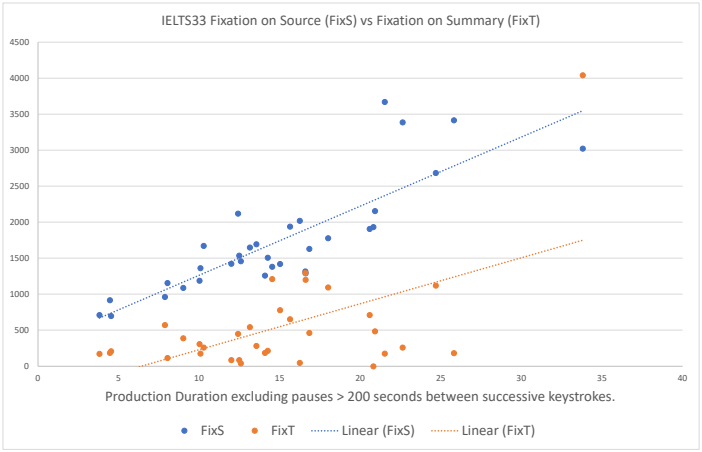


(B) Production duration (x-axis) and Fixation count on Source vs Summary (Y-axis).

FIGURE C.1: Fixation Count on Source vs Summary on CS19 dataset.



(A) Fixation count and total reading time distribution from source and summary in IELTS33 dataset.



(B) Production duration (x-axis) and Fixation count on Source vs Summary (Y-axis).

FIGURE C.2: Fixation Count on Source vs Summary on IELTS33 dataset.

Appendix D

Summarization Evaluation

Part	Abstractive			Extractive		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
P01	0.391	0.099	0.207	NA	NA	NA
P02	0.387	0.043	0.193	NA	NA	NA
P03	0.383	0.061	0.228	0.398	0.127	0.251
P04	0.483	0.153	0.303	0.404	0.100	0.227
P05	0.320	0.087	0.181	0.380	0.064	0.211
P06	0.462	0.121	0.244	0.414	0.171	0.297
P07	0.417	0.116	0.202	0.407	0.077	0.205
P08	0.422	0.107	0.235	0.433	0.169	0.231
P09	0.349	0.083	0.174	0.413	0.109	0.296
P10	0.397	0.068	0.224	0.374	0.091	0.228
P11	0.405	0.092	0.226	0.303	0.059	0.167
P12	0.367	0.013	0.165	0.296	0.104	0.177
P13	0.466	0.196	0.301	0.346	0.127	0.203
P14	0.416	0.138	0.238	0.391	0.082	0.193
P15	0.430	0.123	0.206	0.351	0.089	0.228
P16	0.460	0.135	0.243	0.345	0.106	0.199
P17	0.384	0.115	0.221	0.342	0.090	0.186
P18	0.410	0.044	0.193	0.373	0.110	0.235
P19	0.432	0.077	0.254	0.246	0.065	0.164
P20	0.395	0.082	0.244	0.334	0.096	0.198
P21	0.413	0.090	0.225	0.393	0.113	0.233
P22	0.400	0.118	0.227	0.390	0.071	0.196
P23	0.385	0.081	0.241	0.457	0.138	0.231
P24	0.429	0.103	0.256	0.329	0.069	0.160
P25	0.462	0.135	0.259	NA	NA	NA
P26	0.375	0.075	0.171	0.351	0.068	0.210
P27	0.372	0.075	0.207	0.047	0.024	0.047
P28	0.461	0.118	0.261	NA	NA	NA
P29	0.398	0.069	0.220	NA	NA	NA
P30	0.470	0.133	0.272	0.323	0.076	0.183
Average	0.413	0.099	0.229	0.295	0.080	0.172

TABLE D.1: Students summarization performance on SSG23 dataset.

Text-ID	Part	Abstractive			Extractive		
		Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
U1	P01	0.391	0.099	0.207	NA	NA	NA
U1	P02	0.423	0.012	0.206	NA	NA	NA
U2	P02	0.351	0.075	0.181	NA	NA	NA
U2	P03	0.383	0.061	0.228	0.398	0.127	0.251
U1	P04	0.411	0.087	0.281	0.435	0.131	0.247
U2	P04	0.554	0.220	0.325	0.374	0.070	0.208
U1	P05	0.319	0.078	0.154	0.385	0.089	0.220
U2	P05	0.321	0.096	0.208	0.375	0.039	0.202
U1	P06	0.406	0.105	0.229	0.376	0.113	0.277
U2	P06	0.517	0.136	0.258	0.451	0.229	0.316
U1	P07	0.398	0.095	0.168	0.390	0.053	0.208
U2	P07	0.436	0.138	0.236	0.424	0.102	0.202
U1	P08	0.395	0.081	0.206	0.427	0.160	0.232
U2	P08	0.449	0.133	0.264	0.440	0.178	0.231
U1	P09	0.349	0.083	0.174	0.413	0.109	0.296
U1	P10	0.354	0.025	0.207	0.342	0.056	0.219
U2	P10	0.440	0.111	0.240	0.406	0.126	0.236
U1	P11	0.417	0.096	0.256	0.284	0.029	0.156
U2	P11	0.392	0.089	0.196	0.323	0.089	0.177
U2	P12	0.367	0.013	0.165	0.296	0.104	0.177
U2	P13	0.466	0.196	0.301	0.346	0.127	0.203
U1	P14	0.406	0.127	0.242	0.410	0.073	0.181
U2	P14	0.426	0.149	0.235	0.372	0.091	0.205
U2	P15	0.430	0.123	0.206	0.351	0.089	0.228
U1	P16	0.408	0.083	0.150	0.348	0.096	0.203
U2	P16	0.512	0.188	0.337	0.341	0.116	0.195
U1	P17	0.329	0.055	0.174	0.279	0.092	0.170
U2	P17	0.440	0.174	0.268	0.406	0.088	0.203
U1	P18	0.380	0.059	0.166	0.323	0.103	0.215
U2	P18	0.440	0.029	0.220	0.423	0.118	0.254
U1	P19	0.408	0.090	0.229	0.262	0.076	0.168
U2	P19	0.456	0.064	0.278	0.230	0.054	0.159
U1	P20	0.395	0.082	0.244	0.334	0.096	0.198
U1	P21	0.376	0.102	0.239	0.353	0.097	0.225
U2	P21	0.450	0.077	0.211	0.433	0.129	0.242
U1	P22	0.380	0.131	0.230	0.383	0.033	0.175
U2	P22	0.421	0.104	0.224	0.398	0.110	0.217
U1	P23	0.316	0.059	0.187	0.437	0.110	0.219
U2	P23	0.455	0.103	0.295	0.477	0.165	0.244
U1	P24	0.355	0.072	0.225	0.264	0.083	0.157
U2	P24	0.503	0.133	0.287	0.393	0.055	0.164
U1	P25	0.412	0.094	0.237	NA	NA	NA
U2	P25	0.512	0.176	0.280	NA	NA	NA
U1	P26	0.326	0.071	0.163	0.336	0.065	0.208
U2	P26	0.425	0.079	0.179	0.366	0.071	0.211
U1	P27	0.364	0.053	0.182	0.093	0.048	0.093
U2	P27	0.381	0.097	0.231	NA	NA	NA
U1	P28	0.427	0.074	0.240	NA	NA	NA
U2	P28	0.495	0.163	0.283	NA	NA	NA
U1	P29	0.359	0.044	0.217	NA	NA	NA
U2	P29	0.437	0.094	0.223	NA	NA	NA
U1	P30	0.475	0.153	0.303	0.326	0.084	0.198
U2	P30	0.465	0.113	0.242	0.319	0.068	0.168

TABLE D.2: Student summarization score breakdown for each participants and source texts pair.

	metric	mean	std	min	max
Student					
Abstractive	Rouge-1	0.413	0.054	0.316	0.554
	Rouge-2	0.099	0.045	0.012	0.220
	Rouge-L	0.229	0.044	0.150	0.337
Extractive	Rouge-1	0.361	0.069	0.093	0.477
	Rouge-2	0.096	0.039	0.029	0.229
	Rouge-L	0.208	0.039	0.093	0.316
BART-PT	Rouge-1	0.414	0.012	0.397	0.423
	Rouge-2	0.097	0.014	0.087	0.117
	Rouge-L	0.179	0.018	0.158	0.202
BART-FT	Rouge-1	0.331	0.030	0.291	0.364
	Rouge-2	0.063	0.007	0.054	0.071
	Rouge-L	0.166	0.011	0.151	0.175
DistilBART	Rouge-1	0.338	0.047	0.278	0.393
	Rouge-2	0.082	0.025	0.052	0.113
	Rouge-L	0.208	0.044	0.155	0.262

TABLE D.3: Student and model Rouge score evaluation breakdown from mean, std, min and max

Bibliography

Abnar, Samira and Willem Zuidema (July 2020). "Quantifying Attention Flow in Transformers". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4190–4197. DOI: 10.18653/v1/2020.acl-main.385. URL: https://aclanthology.org/2020.acl-main.385.

Altmann, Gerry T.M, Alan Garnham, and Yvette Dennis (1992). "Avoiding the garden path: Eye movements in context". In: *Journal of Memory and Language* 31.5, pp. 685–712. ISSN: 0749-596X. DOI: https://doi.org/10.1016/0749-596X(92)90035-V. URL: https://www.sciencedirect.com/science/article/pii/0749596X9290035V.

Alvarez-Melis, David and Tommi Jaakkola (Sept. 2017). "A causal framework for explaining the predictions of black-box sequence-to-sequence models". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 412–421. DOI: 10.18653/v1/D17-1042. URL: https://aclanthology.org/D17-1042.

Ancona, Marco et al. (2018). *Towards better understanding of gradient-based attribution methods for Deep Neural Networks*. URL: https://openreview.net/forum?id=Sy21R9JAW.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473*.

Barrett, Maria and Anders Søgaard (Sept. 2015). "Using reading behavior to predict grammatical functions". In: *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*. Lisbon, Portugal: Association for Computational Linguistics, pp. 1–5. DOI: 10.18653/v1/W15-2401. URL: https://aclanthology.org/W15-2401.

Barrett, Maria et al. (Aug. 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)*. Berlin, Germany: Association for Computational Linguistics, pp. 579–584. DOI: 10.18653/v1/P16-2094. URL: https://aclanthology.org/P16-2094.

Barrett, Maria et al. (Oct. 2018a). "Sequence Classification with Human Attention". In: *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Brussels, Belgium: Association for Computational Linguistics, pp. 302–312. DOI: 10.18653/v1/K18-1030. URL: https://aclanthology.org/K18-1030.

Barrett, Maria et al. (June 2018b). "Unsupervised Induction of Linguistic Categories with Records of Reading, Speaking, and Writing". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2028–2038. DOI: 10.18653/v1/N18-1184. URL: https://aclanthology.org/N18-1184.

Belinkov, Yonatan, Sebastian Gehrmann, and Ellie Pavlick (July 2020). "Interpretability and Analysis in Neural NLP". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Online: Association for Computational Linguistics, pp. 1–5. DOI: 10.18653/v1/2020.acl-tutorials.1. URL: https://aclanthology.org/2020.acl-tutorials.1.

Berzak, Yevgeni et al. (2022). "CELER: A 365-Participant Corpus of Eye Movements in L1 and L2 English Reading". In: *Open Mind : Discoveries in Cognitive Science* 6, pp. 41 –50. URL: https://api.semanticscholar.org/CorpusID:250284137.

Binder, Alexander et al. (2016). "Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers". In: *CoRR* abs/1604.00825. arXiv: 1604.00825. URL: http://arxiv.org/abs/1604.00825.

Blohm, Matthias et al. (Oct. 2018). "Comparing Attention-Based Convolutional and Recurrent Neural Networks: Success and Limitations in Machine Reading Comprehension". In: *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Brussels, Belgium: Association for Computational Linguistics, pp. 108–118. DOI: 10.18653/v1/K18-1011. URL: https://aclanthology.org/K18-1011.

Bommasani, Rishi et al. (2023). "The Foundation Model Transparency Index". In: *ArXiv* abs/2310.12941. URL: https://api.semanticscholar.org/CorpusID:264306385.

Brandl, Stephanie and Nora Hollenstein (Nov. 2022). "Every word counts: A multilingual analysis of individual human alignment with model attention". In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online only: Association for Computational Linguistics, pp. 72–77. URL: https://aclanthology.org/2022.aacl-short.10.

Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. arXiv: 2005.14165 [cs.CL].

Carl, Michael (2012a). "The CRITT TPR-DB 1.0: A Database for Empirical Human Translation Process Research". In: *Workshop on Post-Editing Technology and Practice*. San Diego, California, USA: Association for Machine Translation in the Americas. URL: https://aclanthology.org/2012.amta-wptp.1.

— (May 2012b). "Translog-II: a Program for Recording User Activity Data for Empirical Reading and Writing Research". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 4108–

4112. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/614_Paper.pdf.

Carl, Michael and Martin Kay (2012). "Gazing and Typing Activities during Translation: A Comparative Study of Translation Units of Professional and Student Translators". In: *Meta: Translators' Journal* 56, pp. 952–975. URL: https://api.semanticscholar.org/CorpusID:122640017.

Chefer, Hila, Shir Gur, and Lior Wolf (July 2021). "Transformer Interpretability Beyond Attention Visualization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 782–791.

Chen, Peng et al. (Sept. 2017). "Recurrent Attention Network on Memory for Aspect Sentiment Analysis". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 452–461. DOI: 10.18653/v1/D17-1047. URL: https://aclanthology.org/D17-1047.

Cheri, Joe, Abhijit Mishra, and Pushpak Bhattacharyya (Aug. 2016). "Leveraging Annotators' Gaze Behaviour for Coreference Resolution". In: *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*. Berlin: Association for Computational Linguistics, pp. 22–26. DOI: 10.18653/v1/W16-1904. URL: https://aclanthology.org/W16-1904.

Chiang, Wei-Lin et al. (2024). *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*. arXiv: 2403.04132 [cs.AI].

Chowdhery, Aakanksha et al. (2023). "PaLM: Scaling Language Modeling with Pathways". In: *Journal of Machine Learning Research* 24.240, pp. 1–113. URL: http://jmlr.org/papers/v24/22-1144.html.

Cop, Uschi et al. (2017). "Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading". In: *Behavior Research Methods* 49, pp. 602–615.

Culotta, Aron, Andrew McCallum, and Jonathan Betz (June 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*. New York City, USA: Association for Computational Linguistics, pp. 296–303. URL: https://aclanthology.org/N06-1038.

Deng, Shuwen et al. (2023a). "Eyettention: An Attention-based Dual-Sequence Model for Predicting Human Scanpaths during Reading". In: *Proceedings of the ACM on Human-Computer Interaction* 7.ETRA, pp. 1–24.

Deng, Shuwen et al. (Dec. 2023b). "Pre-Trained Language Models Augmented with Synthetic Scanpaths for Natural Language Understanding". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 6500–6507. DOI: 10.18653/v1/2023.emnlp-main.400. URL: https://aclanthology.org/2023.emnlp-main.400.

Devlin, Jacob et al. (June 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). Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

DeYoung, Jay et al. (July 2020). "ERASER: A Benchmark to Evaluate Rationalized NLP Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4443–4458. DOI: 10.18653/v1/2020.acl-main.408. URL: https://aclanthology.org/2020.acl-main.408.

Ding, Shuoyang, Hainan Xu, and Philipp Koehn (Aug. 2019). "Saliency-driven Word Alignment Interpretation for Neural Machine Translation". In: *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Florence, Italy: Association for Computational Linguistics, pp. 1–12. DOI: 10.18653/v1/W19-5201. URL: https://aclanthology.org/W19-5201.

Dong, Yue, Andrei Mircea, and Jackie Chi Kit Cheung (Apr. 2021). "Discourse-Aware Unsupervised Summarization for Long Scientific Documents". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 1089–1102. DOI: 10.18653/v1/2021.eacl-main.93. URL: https://aclanthology.org/2021.eacl-main.93.

Doshi-Velez, Finale and Been Kim (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv: 1702.08608 [stat.ML].

Dziri, Nouha et al. (2023). "Faith and Fate: Limits of Transformers on Compositionality". In: *ArXiv* abs/2305.18654. URL: https://api.semanticscholar.org/CorpusID:258967391.

Eberle, Oliver et al. (May 2022). "Do Transformer Models Show Similar Attention Patterns to Task-Specific Human Gaze?" In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 4295–4309. DOI: 10.18653/v1/2022.acl-long.296. URL: https://aclanthology.org/2022.acl-long.296.

Elhage, Nelson et al. (2021). "A Mathematical Framework for Transformer Circuits". In: *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html. Feng, Shi et al. (2018). "Pathologies of Neural Models Make Interpretations

Difficult". I n: *P roceedings o f t he 2 018 C onference o n E mpirical M ethods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3719–3728. DOI: 10.18653/v1/D18-1407. URL: https://aclanthology.org/D18-1407.

Frazier, Lyn and Keith Rayner (1982). "Making and correcting errors during

sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences". In: *Cognitive Psychology* 14.2, pp. 178 –210. ISSN: 0010-0285. DOI: https://doi.org/10.1016/0010-0285(82)90008-1. URL: http://www.sciencedirect.com/science/article/pii/0010028582900081.

Freitag, Markus and Yaser Al-Onaizan (Aug. 2017). "Beam Search Strategies for Neural Machine Translation". In: *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, pp. 56–60. DOI: 10.18653/v1/W17-3207. URL: https://aclanthology.org/W17-3207.

González-Garduño, Ana Valeria and Anders Søgaard (Sept. 2017). "Using Gaze to Predict Text Readability". In: *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 438–443. DOI: 10.18653/v1/W17-5050. URL: https://aclanthology.org/W17-5050.

Guan, Chaoyu et al. (June 2019). "Towards a Deep and Unified Understanding of Deep Neural Models in NLP". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 2454–2463. URL: https://proceedings.mlr.press/v97/guan19a.html.

Hahn, Michael and Frank Keller (Nov. 2016). "Modeling Human Reading with Neural Attention". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 85–95. DOI: 10.18653/v1/D16-1009. URL: https://aclanthology.org/D16-1009.

— (2018). *Modeling Task Effects in Human Reading with Neural Attention*. arXiv: 1808.00054 [cs.CL].

Hale, John et al. (July 2018). "Finding syntax in human encephalography with beam search". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2727–2736. DOI: 10.18653/v1/P18-1254. URL: https://aclanthology.org/P18-1254.

Hassabis, D. et al. (2017). "Neuroscience-Inspired Artificial Intelligence". In: *Neuron* 95, pp. 245–258.

He, Shilin et al. (Nov. 2019). "Towards Understanding Neural Machine Translation with Word Importance". 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 953–962. DOI: 10.18653/v1/D19-1088. URL: https://aclanthology.org/D19-1088.

Hendrycks, Dan et al. (2020). "Measuring Massive Multitask Language Understanding". In: *ArXiv* abs/2009.03300. URL: https://api.semanticscholar.org/CorpusID:221516475.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Comput.* 9.8, 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

Hollenstein, Nora and Lisa Beinborn (2021a). *Relative Importance in Sentence Processing*. arXiv: 2106.03471 [cs.CL].

Hollenstein, Nora and Lisa Beinborn (Aug. 2021b). "Relative Importance in Sentence Processing". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, pp. 141–150. DOI: 10.18653/v1/2021.acl-short.19. URL: https://aclanthology.org/2021.acl-short.19.

Hollenstein, Nora and Ce Zhang (June 2019). "Entity Recognition at First Sight: Improving NER with Eye Movement Information". 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1–10. DOI: 10.18653/v1/N19-1001. URL: https://aclanthology.org/N19-1001.

Hollenstein, Nora et al. (2018). "ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading". In: *Scientific Data* 5.

Hollenstein, Nora et al. (Nov. 2019a). "CogniVal: A Framework for Cognitive Word Embedding Evaluation". In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, pp. 538–549. DOI: 10.18653/v1/K19-1050. URL: https://aclanthology.org/K19-1050.

Hollenstein, Nora et al. (2019b). *ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation*. arXiv: 1912.00903 [cs.CL].

Honnibal, Matthew and Mark Johnson (Sept. 2015). "An Improved Non-monotonic Transition System for Dependency Parsing". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 1373–1378. DOI: 10.18653/v1/D15-1162. URL: https://aclanthology.org/D15-1162.

Hyönä, Jukka, Raymond A. Bertram, and Alexander Pollatsek (2004). "Are long compound words identified serially via their constituents? Evidence from an eye movement-contingent display change study". In: *Memory & Cognition* 32, pp. 523–532.

Ikhwantri, Fariz et al. (2023). "Looking deep in the eyes: Investigating interpretation methods for neural models on reading tasks using human eye-movement behaviour". In: *Information Processing & Management* 60.2, p. 103195. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2022.103195. URL: https://www.sciencedirect.com/science/article/pii/S0306457322002965.

Jacovi, Alon and Yoav Goldberg (July 2020). "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. URL: https://aclanthology.org/2020.acl-main.386.

Jain, Sarthak and Byron C. Wallace (June 2019). "Attention is not Explanation". 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). Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3543–3556. DOI: 10.18653/v1/N19-1357. URL: https://aclanthology.org/N19-1357.

Jakobsen, Arnt Lykke (1999). "Logging target text production with Translog". In: *Copenhagen studies in language*, pp. 9–20. URL: https://api.semanticscholar.org/CorpusID:64224798.

Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah (July 2019). "What Does BERT Learn about the Structure of Language?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3651–3657. DOI: 10.18653/v1/P19-1356. URL: https://aclanthology.org/P19-1356.

Just, Marcel Adam and Patricia A Carpenter (1976). "Eye fixations and cognitive processes". In: *Cognitive Psychology* 8.4, pp. 441 –480. ISSN: 0010-0285. DOI: https://doi.org/10.1016/0010-0285(76)90015-3. URL: http://www.sciencedirect.com/science/article/pii/0010028576900153.

— (1980). "A theory of reading: from eye fixations to comprehension." In: *Psychological review* 87 4, pp. 329–54.

Kamp, Jonathan, Lisa Beinborn, and Antske Fokkens (May 2024). "The Role of Syntactic Span Preferences in Post-Hoc Explanation Disagreement". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, pp. 16066–16078. URL: https://aclanthology.org/2024.lrec-main.1397.

Kennedy, Andrew B. (2003). "The Dundee Corpus [CD-ROM". In.

Keogh, Eamonn J. et al. (2001). "An Online Algorithm for Segmenting Time Series". In: *Proceedings of the 2001 IEEE International Conference on Data Mining*. ICDM '01. USA: IEEE Computer Society, 289–296. ISBN: 0769511198.

Keskar, Nitish Shirish et al. (2019). "CTRL: A Conditional Transformer Language Model for Controllable Generation". In: *ArXiv* abs/1909.05858. URL: https://api.semanticscholar.org/CorpusID:202573071.

Khurana, Varun et al. (May 2023). "Synthesizing Human Gaze Feedback for Improved NLP Performance". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 1895–1908. URL: https://aclanthology.org/2023.eacl-main.139.

Kim, Yoon (Oct. 2014). "Convolutional Neural Networks for Sentence Classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1746–1751. DOI: 10.3115/v1/D14-1181. URL: https://aclanthology.org/D14-1181.

Kingma, Diederik P. and Jimmy Ba (2017). *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG].

Kullback, S. and R. A. Leibler (1951). "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1, pp. 79–86. ISSN: 00034851. URL: http://www.jstor.org/stable/2236703.

Kumar, Ananya et al. (2022). "Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution". In: *ArXiv* abs/2202.10054. URL: https://api.semanticscholar.org/CorpusID:247011290.

Kuribayashi, Tatsuki, Yohei Oseki, and Timothy Baldwin (June 2024). "Psychometric Predictive Power of Large Language Models". In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 1983–2005. URL: https://aclanthology.org/2024.findings-naacl.129.

Kuribayashi, Tatsuki et al. (Dec. 2022). "Context Limitations Make Neural Language Models More Human-Like". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 10421–10436. DOI: 10.18653/v1/2022.emnlp-main.712. URL: https://aclanthology.org/2022.emnlp-main.712.

Lewis, Mike et al. (July 2020). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703.

Li, Jing, Aixin Sun, and Shafiq Joty (July 2018). "SegBot: A Generic Neural Text Segmentation Model with Pointer Network". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, pp. 4166–4172. DOI: 10.24963/ijcai.2018/579. URL: https://doi.org/10.24963/ijcai.2018/579.

Lin, Chin-Yew (July 2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: https://aclanthology.org/W04-1013.

Lipton, Zachary C. (June 2018). "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery." In: *Queue* 16.3, 31–57. ISSN: 1542-7730. DOI: 10.1145/3236386.3241340. URL: https://doi.org/10.1145/3236386.3241340.

Liu, Yang and Mirella Lapata (Nov. 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 3730–3740. DOI: 10.18653/v1/D19-1387. URL: https://aclanthology.org/D19-1387.

Lu, Yu et al. (2022). "Attention Analysis and Calibration for Transformer in Natural Language Generation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30, pp. 1927–1938. URL: https://api.semanticscholar.org/CorpusID:249563587.

Luong, Thang, Hieu Pham, and Christopher D. Manning (Sept. 2015). "Effective Approaches to Attention-based Neural Machine Translation". In:

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, pp. 1412–1421. DOI: 10.18653/v1/D15-1166. URL: https://aclanthology.org/D15-1166.

Maharaj, Kishan et al. (Dec. 2023). "Eyes Show the Way: Modelling Gaze Behaviour for Hallucination Detection". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 11424–11438. DOI: 10.18653/v1/2023.findings-emnlp.764. URL: https://aclanthology.org/2023.findings-emnlp.764.

Maki, Ryosuke, Hitoshi Nishikawa, and Takenobu Tokunaga (Dec. 2016). "Parameter estimation of Japanese predicate argument structure analysis model using eye gaze information". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 2861–2869. URL: https://aclanthology.org/C16-1269.

Malmaud, Jonathan, Roger Levy, and Yevgeni Berzak (Nov. 2020). "Bridging Information-Seeking Human Gaze and Machine Reading Comprehension". In: *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, pp. 142–152. DOI: 10.18653/v1/2020.conll-1.11. URL: https://aclanthology.org/2020.conll-1.11.

Manning, Christopher D., Kevin Clark, and John Hewitt (2020). "Emergent linguistic structure in artificial neural networks trained by self-supervision". In: *Proceedings of the National Academy of Sciences*. Vol. 117, pp. 30046–30054.

Marcu, Daniel (2000). *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA, USA: MIT Press. ISBN: 0262133725.

Marcu, Daniel and Abdessamad Echihabi (July 2002). "An Unsupervised Approach to Recognizing Discourse Relations". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 368–375. DOI: 10.3115/1073083.1073145. URL: https://aclanthology.org/P02-1047.

Mathias, Sandeep et al. (July 2018). "Eyes are the Windows to the Soul: Predicting the Rating of Text Quality Using Gaze Behaviour". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2352–2362. DOI: 10.18653/v1/P18-1219. URL: https://aclanthology.org/P18-1219.

Mishra, Abhijit, Michael Carl, and Pushpak Bhattacharyya (Dec. 2012). "A heuristic-based approach for systematic error correction of gaze data for reading". In: *Proceedings of the First Workshop on Eye-tracking and Natural Language Processing*. Mumbai, India: The COLING 2012 Organizing Committee, pp. 71–80. URL: https://aclanthology.org/W12-4906.

Mishra, Abhijit, Kuntal Dey, and Pushpak Bhattacharyya (July 2017). "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)*. Vancouver, Canada: Association for Computational Linguistics, pp. 377–387. DOI: 10.18653/v1/P17-1035. URL: https://aclanthology.org/P17-1035.

Mishra, Abhijit, Diptesh Kanojia, and Pushpak Bhattacharyya (2016). "Predicting Readers' Sarcasm Understandability by Modeling Gaze Behavior". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 30.1. DOI: 10.1609/aaai.v30i1.9884. URL: https://ojs.aaai.org/index.php/AAAI/article/view/9884.

Myers, Leann and Maria J. Sirois (2006). "Spearman Correlation Coefficients, Differences between". In: *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Ltd. ISBN: 9780471667193. DOI: https://doi.org/10.1002/0471667196.ess5050.pub2. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471667196.ess5050.pub2. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471667196.ess5050.pub2.

Nallapati, Ramesh et al. (Aug. 2016). "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond". In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, pp. 280–290. DOI: 10.18653/v1/K16-1028. URL: https://aclanthology.org/K16-1028.

Narayan, Shashi, 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*. Brussels, Belgium: Association for Computational Linguistics, pp. 1797–1807. DOI: 10.18653/v1/D18-1206. URL: https://aclanthology.org/D18-1206.

Naseer, Muhammad Muzammal et al. (2021). "Intriguing Properties of Vision Transformers". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 23296–23308. URL: https://proceedings.neurips.cc/paper/2021/file/c404a5adbf90e09631678b13b05d9d7a-Paper.pdf.

Nilsson, Mattias and Joakim Nivre (June 2009). "Learning Where to Look: Modeling Eye Movements in Reading". In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*. Boulder, Colorado: Association for Computational Linguistics, pp. 93–101. URL: https://aclanthology.org/W09-1113.

Oba, Miyu et al. (July 2023). "Second Language Acquisition of Neural Language Models". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 13557–13572. DOI: 10.18653/v1/2023.findings-acl.856. URL: https://aclanthology.org/2023.findings-acl.856.

Oh, Byung-Doh and William Schuler (2023). "Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times?" In: *Transactions of the Association for Computational Linguistics*

11, pp. 336–350. DOI: 10.1162/tacl_a_00548. URL: https://aclanthology.org/2023.tacl-1.20.

Oh, Byung-Doh, Shisen Yue, and William Schuler (Mar. 2024). "Frequency Explains the Inverse Correlation of Large Language Models' Size, Training Data Amount, and Surprisal's Fit to Reading Times". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, pp. 2644–2663. URL: https://aclanthology.org/2024.eacl-long.162.

Olsson, Catherine et al. (2022). "In-context Learning and Induction Heads". In: *Transformer Circuits Thread*. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Parikh, Ankur et al. (Nov. 2016). "A Decomposable Attention Model for Natural Language Inference". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2249–2255. DOI: 10.18653/v1/D16-1244. URL: https://aclanthology.org/D16-1244.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

Poel, Liam van der, Ryan Cotterell, and Clara Meister (Dec. 2022). "Mutual Information Alleviates Hallucinations in Abstractive Summarization". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5956–5965. DOI: 10.18653/v1/2022.emnlp-main.399. URL: https://aclanthology.org/2022.emnlp-main.399.

Poerner, Nina, Hinrich Schütze, and Benjamin Roth (July 2018). "Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 340–350. DOI: 10.18653/v1/P18-1032. URL: https://aclanthology.org/P18-1032.

Poole, A and Linden Ball (Jan. 2006). "Eye tracking in human-computer interaction and usability research: Current status and future prospects". In: *Encyclopedia of Human Computer Interaction*, pp. 211–219.

Pouw, Charlotte, Nora Hollenstein, and Lisa Beinborn (May 2023). "Cross-Lingual Transfer of Cognitive Processing Complexity". In: *Findings of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 655–669. URL: https://aclanthology.org/2023.findings-eacl.49.

Radford, Alec et al. (2019). "Language Models are Unsupervised Multitask Learners". In: URL: https://api.semanticscholar.org/CorpusID:160025533.

Rajpurkar, Pranav et al. (Nov. 2016). "SQuAD: 100,000+ Questions for Machine Comprehension of Text". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2383–2392. DOI: 10.18653/v1/D16-1264. URL: https://aclanthology.org/D16-1264.

Rayner, Keith (1998). "Eye movements in reading and information processing: 20 years of research." In: *Psychological bulletin* 124 3, pp. 372–422.

— (2009). "Eye movements and attention in reading, scene perception, and visual search". In: *The Quarterly Journal of Experimental Psychology* 62.8, pp. 1457–1506.

Reichle, Erik D., Keith Rayner, and Alexander Pollatsek (2003). "The E-Z Reader model of eye-movement control in reading: Comparisons to other models". In: *Behavioral and Brain Sciences* 26.4, 445–476. DOI: 10.1017/S0140525X03000104.

Reichle, Erik D. et al. (1998). "Toward a model of eye movement control in reading". In: *Psychological Review* 105.1, pp. 125–157.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.

Richardson, Daniel C., Rick Dale, and Michael J. Spivey (2007). "Eye movements in language and cognition: A brief introduction". In: *Methods in Cognitive Linguistics*. Ed. by Monica Gonzalez-Marquez et al. John Benjamins., pp. 323–344.

Rodeghero, Paige and Collin McMillan (2015). "An Empirical Study on the Patterns of Eye Movement during Summarization Tasks". In: *2015 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pp. 1–10. DOI: 10.1109/ESEM.2015.7321188.

Sahoo, Debasish and Michael Carl (Aug. 2019). "Lexical Representation & Retrieval on Monolingual Interpretative text production". In: *Proceedings of the Second MEMENTO workshop on Modelling Parameters of Cognitive Effort in Translation Production*. Dublin, Ireland: European Association for Machine Translation, pp. 14–16. URL: https://aclanthology.org/W19-7007.

Sawaki, Yasuyo (2020). "Developing Summary Content Scoring Criteria for University L2 Writing Instruction in Japan". In: URL: https://api.semanticscholar.org/CorpusID:229654958.

Schou, Lasse, Barbara Dragsted, and Michael Carl (2009). "Ten years of Translog". In: *Copenhagen studies in language*, pp. 37–48. URL: https://api.semanticscholar.org/CorpusID:63770787.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (Aug. 2016). "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162.

Seo, Minjoon et al. (2017). "Bidirectional Attention Flow for Machine Comprehension". In: *ArXiv* abs/1611.01603.

Serrano, Sofia and Noah A. Smith (July 2019). "Is Attention Interpretable?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2931–2951. DOI: 10.18653/v1/P19-1282. URL: https://aclanthology.org/P19-1282.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017). "Learning Important Features through Propagating Activation Differences". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 3145–3153.

Siegelman, Noam et al. (2022). "Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO)". In: *Behavior Research Methods* 54, pp. 2843 –2863. URL: https://api.semanticscholar.org/CorpusID:246488575.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2013). *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. arXiv: 1312.6034 [cs.CV].

Socher, Richard et al. (Oct. 2013). "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1631–1642. URL: https://aclanthology.org/D13-1170.

Sood, Ekta et al. (2020a). "Improving Natural Language Processing Tasks with Human Gaze-Guided Neural Attention". In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–15. URL: https://proceedings.neurips.cc/paper/2020/hash/460191c72f67e90150a093b4585e7eb4-Abstract.html.

Sood, Ekta et al. (Nov. 2020b). "Interpreting Attention Models with Human Visual Attention in Machine Reading Comprehension". In: *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, pp. 12–25. DOI: 10.18653/v1/2020.conll-1.2. URL: https://aclanthology.org/2020.conll-1.2.

Stiennon, Nisan et al. (2020). "Learning to Summarize from Human Feedback". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 3319–3328.

Tang, Joël, Marina Fomicheva, and Lucia Specia (2023). *Reducing Hallucinations in Neural Machine Translation with Feature Attribution*. arXiv: 2211.09878 [cs.CL].

Tapaswi, Makarand et al. (2016). "MovieQA: Understanding Stories in Movies through Question-Answering". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4631–4640.

Tardy, Paul et al. (May 2020). "Align then Summarize: Automatic Alignment Methods for Summarization Corpus Creation". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 6718–6724. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.829.

Tokunaga, Takenobu, Hitoshi Nishikawa, and Tomoya Iwakura (Sept. 2017). "An Eye-tracking Study of Named Entity Annotation". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. Varna, Bulgaria: INCOMA Ltd., pp. 758–764. DOI: 10.26615/978-954-452-049-6_097. URL: https://doi.org/10.26615/978-954-452-049-6_097.

Touvron, Hugo et al. (2023). "LLaMA: Open and Efficient Foundation Language Models". In: *ArXiv* abs/2302.13971. URL: https://api.semanticscholar.org/CorpusID:257219404.

Vafa, Keyon et al. (Nov. 2021). "Rationales for Sequential Predictions". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 10314–10332. DOI: 10.18653/v1/2021.emnlp-main.807. URL: https://aclanthology.org/2021.emnlp-main.807.

Vashishth, Shikhar et al. (2019). *Attention Interpretability Across NLP Tasks*. arXiv: 1909.11218 [cs.CL].

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Vig, Jesse (July 2019). "A Multiscale Visualization of Attention in the Transformer Model". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florence, Italy: Association for Computational Linguistics, pp. 37–42. DOI: 10.18653/v1/P19-3007. URL: https://aclanthology.org/P19-3007.

Voita, Elena, Rico Sennrich, and Ivan Titov (Aug. 2021). "Analyzing the Source and Target Contributions to Predictions in Neural Machine Translation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 1126–1140. DOI: 10.18653/v1/2021.acl-long.91. URL: https://aclanthology.org/2021.acl-long.91.

Wang, Alex et al. (2019). "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

Wiegreffe, Sarah and Yuval Pinter (Nov. 2019). "Attention is not not Explanation". 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)*. Hong Kong, China: Association for

Computational Linguistics, pp. 11–20. DOI: 10.18653/v1/D19-1002. URL: https://aclanthology.org/D19-1002.

Wiseman, Sam and Alexander M. Rush (Nov. 2016). "Sequence-to-Sequence Learning as Beam-Search Optimization". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1296–1306. DOI: 10.18653/v1/D16-1137. URL: https://aclanthology.org/D16-1137.

Wolf, Thomas et al. (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6.

Xiong, Caiming, Victor Zhong, and Richard Socher (2016). *Dynamic Coattention Networks For Question Answering*. arXiv: 1611.01604 [cs.CL].

Xu, Jiacheng and Greg Durrett (Aug. 2021). "Dissecting Generation Modes for Abstractive Summarization Models via Ablation and Attribution". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 6925–6940. DOI: 10.18653/v1/2021.acl-long.539. URL: https://aclanthology.org/2021.acl-long.539.

Xu, Jiacheng et al. (July 2020). "Discourse-Aware Neural Extractive Text Summarization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5021–5031. DOI: 10.18653/v1/2020.acl-main.451. URL: https://aclanthology.org/2020.acl-main.451.

Xu, Weijia et al. (2023). "Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection". In: *Transactions of the Association for Computational Linguistics* 11, pp. 546–564. DOI: 10.1162/tacl_a_00563. URL: https://aclanthology.org/2023.tacl-1.32.

Yang, Zhilin et al. (2019). *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. DOI: 10.48550/ARXIV.1906.08237. URL: https://arxiv.org/abs/1906.08237.

Yu, Adams Wei et al. (2018). "QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension". In: *ArXiv* abs/1804.09541.

Zeiler, Matthew D and Rob Fergus (2013). *Visualizing and Understanding Convolutional Networks*. DOI: 10.48550/ARXIV.1311.2901. URL: https://arxiv.org/abs/1311.2901.

Zeiler, Matthew D. and Rob Fergus (2014). "Visualizing and Understanding Convolutional Networks". In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Cham: Springer International Publishing, pp. 818–833. ISBN: 978-3-319-10590-1.

Zelinsky, Gregory et al. (2006). "The Role of Top-down and Bottom-up Processes in Guiding Eye Movements during Visual Search". In: *Advances in Neural Information Processing Systems*. Ed. by Y. Weiss, B. Schölkopf, and J. Platt. Vol. 18. MIT Press. URL: https://proceedings.neurips.cc/paper/2005/file/564645fbd0332f066cbd9d083ddd077c-Paper.pdf.