Interactive Attention Model Explorer for Natural Language Processing Tasks with Unbalanced Data Sizes

Zhihang Dong* Tongshuang Wu† Sicheng Song Mingrui Zhang

University of Washington

Abstract

Conventional attention visualization tools compromise either the readability or the information conveyed when documents are lengthy, especially when these documents have imbalanced sizes. Our work strives toward a more intuitive visualization for a subset of Natural Language Processing tasks, where attention is mapped between documents with imbalanced sizes. We extend the flow map visualization to enhance the readability of the attention-augmented documents. Through interaction, our design enables semantic filtering that helps users prioritize important tokens and meaningful matching for an in-depth exploration. Case studies and informal user studies in machine comprehension prove that our visualization effectively helps users gain initial understandings about what their models are "paying attention to." We discuss how the work can be extended to other domains, as well as being plugged into more end-to-end systems for model error analysis.

Index Terms: Human-centered computing—Visualization—Visualization techniques—Heatmaps; Human-centered computing—Visualization—Interaction design process and methods

1 Introduction

The attention mechanism [1] has been widely used in natural language processing (NLP) tasks including sentiment analysis [7, 8], question answering [19], and text summarization [18]. It has been the most intuitive intermediate delivery from neural network models up-to-date: by exporting vectors of attention weights corresponding to the encoding units in encoder-decoder models, attention approximates what the models "are looking at", hence helps debugging common problems like repetition and copying in black-box models.

Attention visualizations have effectively helped researchers to verify alignments in machine translation models. Unfortunately, most representative visualizations (e.g., attention heatmaps [1] and flow maps [16]) suffer from visual clutter issues when tasks contain long documents or unbalanced document sizes. Moreover, static attention visualization usually naively conveys all of the attention entries while omitting important linguistic features that might be, in fact, more important than attention entries. For example, a person-name matching is usually more meaningful than a preposition matching. Unfortunately, albeit the demonstrated importance, these aspects are not easily represented with the current state of the art.

To address such issues, we design an interactive attention visualization for NLP tasks where attention model actions/units are mapped between documents with imbalanced sizes. Such cases are presented in a notable subset of NLP applications, such as question answering (whose inputs are long paragraphs and short questions), text summarization (whose models take in original passages and output shortened summaries). Specifically, we extend the flow map to compactly display the long documents in a fixed screen size. We closely link the compressed flow map to the original documents so that readability is effectively recovered. Through interaction, we also enable some semantic filtering to help users prioritize (1) anchor tokens (e.g., only display named entities) or (2) meaningful matchings (matched on exact lemma, on POS tags, or on random tokens) so that in-depth exploration is made possible.

Requirements	text heatmap	2D heatmap	Flow
Readable	✓	✗	✓
Scalable	✓	✗	✗
Overview	✓	✗	✗
Align.	✗	✓	✓

Table 1: Four requirements for effective attention visualizations (R1-R4). Existing visualizations all have limitations. Meanwhile, our proposed method satisfies all the requirements.

We demonstrate our visualization with three representative examples in the context of question-answering (QA) tasks. An informal user study shows that our visualization effectively helps users gain initial understandings about what their models are "paying attention to". We conclude our study with extensive discussions on how the visual design can be extended for other domains and analysis tasks.

2 Related Work

2.1 Visualization for NLP Tasks

There have been a growing number of NLP works in recent years thanks to the development of areas like deep learning. However, with most models being black boxes [14], our understandings on such enhancements have yet to be sufficient. In order to help analysts evaluate their models and data, various studies (e.g., [5]) try to bundle interactive visualization with text mining and summarization techniques. To date, visualizations have been proved helpful in pinpointing ambiguity and incompleteness [3], understanding and identifying potential issues in neural models [7], analyzing model outcomes [6] and providing interactive visualization to neural network weights [17].

Existing studies typically try to "open the black box" in two methods. On the one hand, various studies have tried to visualizing the activation of neurons, such studies reveal the hidden state dynamics [10, 20]. While such work dives deep into the recurrent neural network architecture, it is usually constrained by the specific model design, with limited generalizability to other model structures. On the other hand, more studies focus on visualizing the model attentions. With the attention being a commonly available layer in most state-of-the-art NLP models, attention visualization tend to be more applicable in various use cases. Our work also falls under this domain, and we will discuss attention visualization in the next section.

2.2 Challenges in Attention Visualization

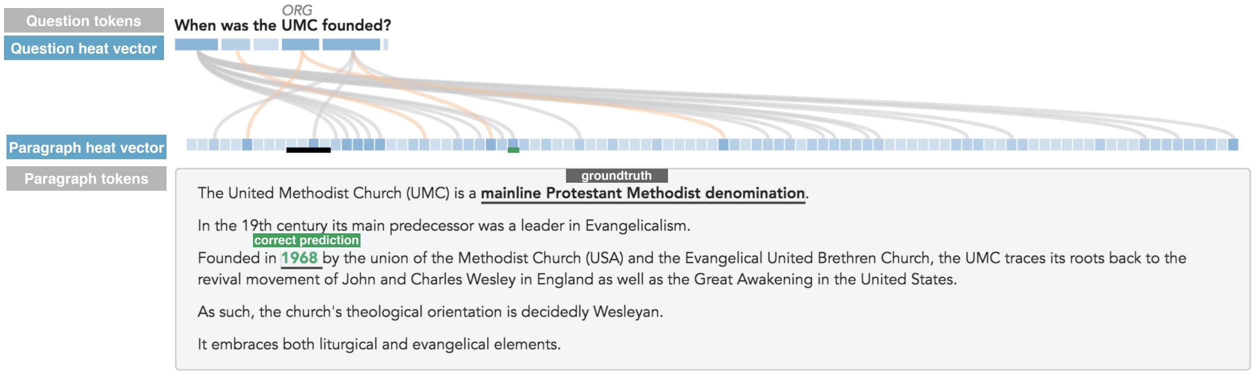
In this section, we discuss the current state of the art in the visualization of attention model. In Table 1, we summarize these methods into three categories, and review them based on the four design requirements:

*zdong@uw.edu

†wtshuang@cs.washington.edu

Zhihang Dong and Tongshuang Wu made equal contribution to the paper.

2020 IEEE Pacific Visualization Symposium (PacificVis)
978-1-7281-5697-2/20/$31.00 ©2020 IEEE



The United Methodist Church (UMC) is a mainline Protestant Methodist denomination.

In the 19th century its main predecessor was a leader in Evangelicalism.

Founded in 1968 by the union of the Methodist Church (USA) and the Evangelical United Brethren Church, the UMC traces its roots back to the revival movement of John and Charles Wesley in England as well as the Great Awakening in the United States.

As such, the church's theological orientation is decidedly Wesleyan.

It embraces both liturgical and evangelical elements.

Figure 1: Attention overview, with readable layouts for the question and paragraph, and a flow map encoding the token importance and alignments.

R1 **Preserves the readability of the text.** The source texts should be readable along with the visualization, such that analysts can intuitively align the semantics and the model's attention.

R2 **Scales to long documents.** The visualization should be compatible to long textual paragraphs or documents (as frequently seen in question answering or text summarization tasks.)

R3 **Provides an overview.** The visualization should convey, overall, which sentences or short phrases a model is attending to.

R4 **Aligns imbalanced text sources.** For tasks with multiple text sources, it is important to highlight how the model is attending to pairs of related tokens.

Text Heatmap [9] colors the background of each text token on a gradient scale from white to a specific color, with deeper colors representing more significant values. Tasks with only one input document (e.g., sentiment analysis [7,8]) usually use in-text highlighting (or textual heatmap) that intuitively enables the within-paragraph comparison. This visualization technique can be very effective in single-document tasks such as sentiment evaluation. However, its lack-of-alignment is less effective for tasks with multiple text sources (✗ R1). Especially when the documents are of different lengths, heatmap can be very difficult to comprehend due to its lack of indication on multi-text correspondence.

2D Heatmap [1] provides a correlation matrix of different tokens within a finite space. It is easy to cross-compare the correlation of two tokens compared to that of another. However, a disadvantage with such 2D matrices is that they take too much space (✗ R2) With tasks involving large input or output (e.g. a hundred or more tokens) the size of the heat-map quickly gets out of hand. Exhaustive scrolling greatly hinders analysts from getting a quick overview (✗ R3), and decreases the effectiveness of a visualization with respect to analysis tasks. At the same time, such methods make them very difficult to read (✗ R1). Source text is rarely in a token-per-line format, as is presented in a 2D heatmap. In such cases, we would likely lose insightful information that could be drawn from analyzing the original structure of the text.

Flow Map [16] encodes attention weights between two related tokens with thick edges pairing the tokens. As demonstrated by [11], flow map provide dynamic visualization of interactions and details for individual tokens. Flow maps have been demonstrated to effectively visualize a large variety of data [4], particularly those with high dimensions [2]. Flow map, on the other hand, unfortunately, suffers from the problem of poor alignment scalability (✗ R2), which means it usually requires a matched input-output size [12]. Furthermore, with the links between tokens thickly overlapping with each other, many of the earlier attempts on flow map for visualization lack a good overview (✗ R3) and enough high level semantic grouping.

Inspired by acquiring the advantages of both text heat map and the strengths of flow map techniques, we build a flexible visualization model with design consideration of all above highlights included. We shall discuss our model design next section in greater detail.

3 Design Consideration

To design a visual system exploring the mechanism of attention layers with high flexibility, interpretability, readability, we aim to achieve three overarching goals: compact overview, selective token views, linguistic match highlights. Such three aspects are especially important to our design as they conquer the difficulties mentioned above. Therefore, it should (1) provide a compact overview so users sense the matching of tokens with questions; (2) select relevant tokens concerned by the model with user-defined part-of-questions; (3) highlight matches with accuracy measurement that signifies its linguistic importance given a context. In the next few paragraphs, we discuss each of the following methodological concerns in greater details.

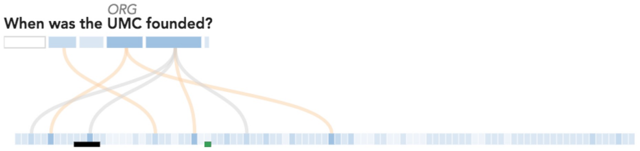
Compact Overview To provide a global context, we highlight the within-document attention distribution and the between-document alignments to understand (1) what tokens are matched by toggling the selection and (2) how attention are "distributed by our model" — is it dense in some regions, or just evenly distributed across the whole document? Such within-document distribution supports explorations on how our model finds the answer, while the between-document alignment reveals the mechanism of the model .

Selective Token Views Unnecessary tokens (e.g., prepositions) confuses readers because the relevant and matched tokens may not be of the same importance. Our design allows users to toggle targeted tokens, hence viewing the distribution of attention on only one specific token or to filter out unnecessary ones, which eases the visual clutter issues and improves focuses.

Linguistic Match Highlights Abstracting raw texts into linguistically meaningful tags helps users to locate semantically similar tokens. For instance, in the automatic question-answering context, let the ground-truth answer be a ”person-name”, highlighting some other "person" tokens could have easily helped the spot potential answers. This mechanism deduces helpful visual cues by mimicking the judgment of deep natural language processing models on their assigned tasks so that models are much more interpretable.

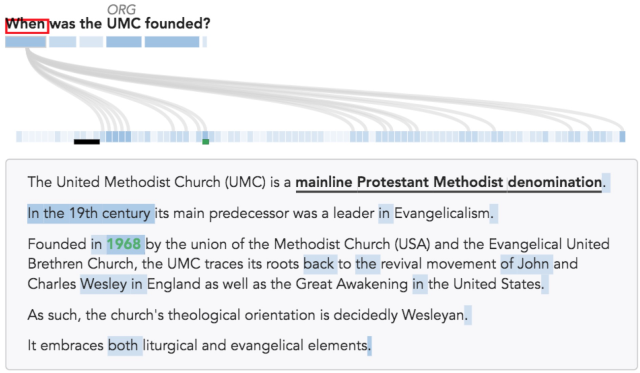
4 Visual Design

For reader's convenience, the following discussion will be based on a question-answering task. The data come from the SQuAD [13] project, a reading comprehension data set consisting of questions posed by crowdworkers on a set of Wikipedia articles. SQuAD includes more than 100,000 instances, each contains a tuple of question, (background) paragraph and their corresponding groundtruth answer — a segment of text or span from the corresponding passage.



When was the UMC founded?

Figure 2: Disable the attentions on “when” to get a clearer flow map.



When was the UMC founded?

The United Methodist Church (UMC) is a **mainline Protestant Methodist denomination**.

In the 19th century its main predecessor was a leader in Evangelicalism.

Founded in 1968 by the union of the Methodist Church (USA) and the Evangelical United Brethren Church, the UMC traces its roots back to the revival movement of John and Charles Wesley in England as well as the Great Awakening in the United States.

As such, the church's theological orientation is decidedly Wesleyan.

It embraces both liturgical and evangelical elements.

Figure 3: Pinpointedly view attentions on “when” to see matched dates and preprositions.

Our analytical model is the bi-directional attention model [19] based on the calculated pairwise token attention weight between questions and paragraphs.

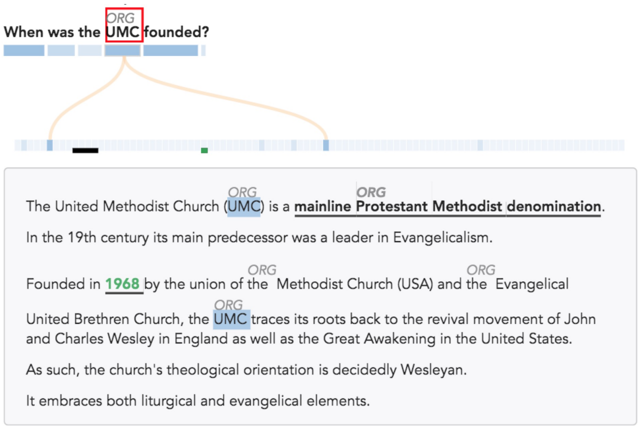
4.1 Overview

Our design takes advantages of the strengths of flow map and the heatmap – both are commonly used in NLP tasks to lower the learning curve. As shown in Fig. 1, we present the layout of the question-and-paragraph-pair in a naturally readable way. Answers are colored and underlined inside the paragraphs, with black being the groundtruth, red being the incorrect prediction from the model, and green the correct ones. We project each question and paragraph token to a single rectangle, and employ flow map to show the high-level importance distribution and alignment structure. The question rectangles are aligned with the corresponding token, and the ones for the paragraph tokens are compressed into one row. These rectangles are colored based on the best matched attention score. This can be seen as a heatmap compressed on either the question token dimension or the paragraph token dimension. For instance, to fill the paragraph rectangles, we take the max overall all the question tokens for each paragraph token. These rows therefore encodes the overall token importance distribution. For the alignment, if the attention score between one question token and one paragraph token passes the threshold (in empirical case, 0.5), we link the corresponding two rectangles for the purpose of encoding. In addition, exact matches (e.g., “UMC” in Fig. 4) are colored in orange, and grey links implies potential relations between two tokens by other attention units (e.g., question token “when” and paragraph token denoting dates like “1968” in Fig. 3).

4.2 Implementation Details

The front-end of this visual system is written using Node JS (npm version 8.9.0 and 5.8.0) and Typescript. The back-end of this visual system is written using Python version 3.6.0. To do production build and upload to the code-share repositories, users only need to run a short script.

The build time of our visual system is proportional to the complication of the natural language processing task, but the build of this visual system is competitively fast and scalable to larger text documents. The user interface instantly reacts to each user query because such query has been already computed beforehand.



When was the UMC founded?

The United Methodist Church (UMC) is a **mainline Protestant Methodist denomination**.

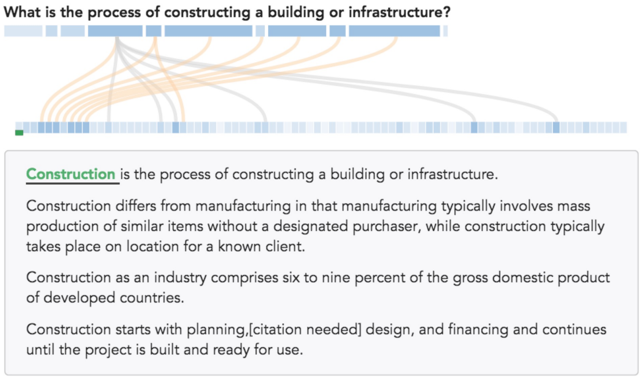
In the 19th century its main predecessor was a leader in Evangelicalism.

Founded in 1968 by the union of the Methodist Church (USA) and the Evangelical United Brethren Church, the UMC traces its roots back to the revival movement of John and Charles Wesley in England as well as the Great Awakening in the United States.

As such, the church's theological orientation is decidedly Wesleyan.

It embraces both liturgical and evangelical elements.

Figure 4: Triggering the named entity annotation “ORG” in the paragraph when “UMC” is selected.


What is the process of constructing a building or infrastructure?

Construction is the process of constructing a building or infrastructure.

Construction differs from manufacturing in that manufacturing typically involves mass production of similar items without a designated purchaser, while construction typically takes place on location for a known client.

Construction as an industry comprises six to nine percent of the gross domestic product of developed countries.

Construction starts with planning,[citation needed] design, and financing and continues until the project is built and ready for use.

Figure 5: A case where the model’s prediction exactly match with the groundtruth text span. From the orange links, we can easily see that the question has high text overlap with the first sentence in the paragraph, which contains the groundtruth answer.

4.3 Interactions

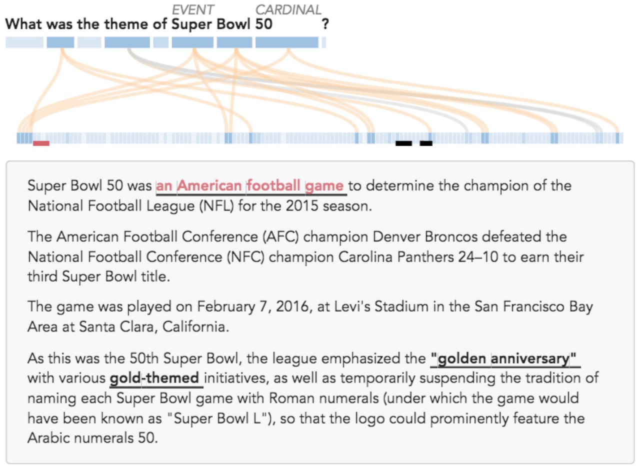
We link flow map and heatmap with interactions and semantic filtering features to highlight user inquiries on useful tokens. Users could pinpoint and display attentions for certain tokens either by either diving into only one question token or by toggling certain unnecessary tokens. For instance, in Fig. 3), hovering on the token “when” filters all the links connected to “when”, re-colors the paragraph rectangles, and highlights the matched tokens in the paragraph. Here, as “when” is matched to multiple paragraph tokens hence creates visual clutters, we hide their arrival at Fig. 2) so that users could hide less informative tokens like “was”.

Moreover, we also highlight matches with accuracy measurement that signify its linguistic importance given a context. In Fig. 4, hovering on “UMC”, a named entity for “ORG” (organization), triggers all the “ORG” tokens in the paragraph.

5 Evaluations

5.1 Use Case

We demonstrate the usefulness of our visualization design with two contrasting use cases.



Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season.

The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title.

The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

As this was the 50th Super Bowl, the league emphasized the **"golden anniversary"** with various **gold-themed** initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

Figure 6: A case where attentions are dragged to two far ends and confuses the model. The matched tokens are very distant to the groundtruth spans (highlighted in black), which could cause the misprediction.

Fig. 5 provides an example where we have exact, correct predicted matches that correspond to the groundtruth. The tokens following the correct answer are highly responsive to the tokens after the focal token of the example question.

In contrast, Fig. 6 illuminates the scenario when our model renders a prediction that the answer produced is likely incorrect: the matched tokens are very distant to the paragraph token serving as the predicted answer. Given its more distant characteristics, our model reaches the conclusion that its prediction is likely incorrect. Users are able to observe this decision-making process by considering the distance kept between the two ’clusters’ of links.

5.2 Informal User Study

Description To examine the effectiveness of our visual design, we conducted a user study on the college campus where students across different majors have exposure to ($n = 61$). In each case, participants were recruited by his/her own will and there was no reward to participate. In each participation session, we briefly introduced the nature of a question-answering task in natural language processing and the dataset to the participant regardless of his/her prior knowledge in natural language processing or artificial intelligence in general. All participants were asked their previous experience in deep learning visualization (such as Tensor Board) so that hindsight bias is controlled. The participants were shown the four visualizations (text heatmap, 2D heatmap, flow map and our visualization design) one by one, and were asked to evaluate them promptly after each design demonstration. To avoid biases, we randomly assigned different orders of demonstration of four designs to participants, so they did not know which one was our design. A complete evaluation of all four designs was mandatory to be counted into our evaluation dataset. Specifically, we ask each of the participant to evaluate each of the four designs on three criteria using a scale of 1 (very poor) to 5 (very good) focused on readability, overview and alignment, which align to our three overarching goals: compact selective token views, overview and linguistic match highlights:

- How readable the current design was in demonstrating relationships within the document and between the paired documents?
- How well the current design did in reducing confusion on finding the relevant and matched languages, staying away from visual clutter issues and helping you focus?
- How well did the visual cues in this design mimic the judgment of deep learning models on the QA tasks such that models are much more interpretable?

Results Table 2 is the evaluation of the three criteria on text heatmap, 2D heatmap, flow map and our design based on 61 college students. The first line is the mean score and the second line is the standard deviation in the parentheses.

Task	Text map	2D map	Flow map	Our Design
Readability	2.739	2.087	3.391	**3.884**
	(1.196)	(1.269)	(1.032)	(0.738)
Overview	2.304	3.333	2.913	**4.087**
	(1.478)	(1.196)	(1.531)	(0.935)
Alignment	1.565	1.275	2.927	**3.261**
	(0.717)	(0.450)	(1.129)	(1.291)

Table 2: User Study Evaluation Results.

In each of the three criteria, our design outperforms the three widely adopted design. To test the statistical significance of our results, we run a ”toughest competitor” t-test against the second best performers in the user evaluation. The t-test shows that our design did significantly better in readability task ($t = -3.2254$, d.f. = 123.16, p-value = 0.002) and overview task ($t = -4.1223$, d.f. = 128.51, p-value < 0.001). For the alignment task, the t-test shows a marginal level of difference ($t = -1.6148$, d.f. = 133.62, p-value = 0.100). We then investigated the association between the scoring difference of rating in different design frameworks and the levels of claimed knowledge on natural language processing and artificial intelligence using the Pearson’s r-correlation. We did not find any beyond trivial association ($r = -0.043$). We conclude the study by an open-ended question asking them to comment on the concerns they have on our design, which shall be discussed in our Discussion.

6 Discussion

Our study provides an interactive, intuitive and easily interpretable visualizer of attention model that combines multiple conventional visualization methods. By compactly displaying the attentions for long documents, our visualization becomes more scalable, helps provide overviews even when long documents are present. With the heatmap and the flow map integrated, we achieve alignments between imbalanced text sources. Through interaction, in-depth explorations on targeted tokens helps analysts flexibly explore their models’ behaviors-of-interest. Case studies and informal user studies in machine comprehension prove that our visualization effectively helps users gain initial understandings about what their models are “paying attention to.” Below, we envision some practical use cases of our proposed visualization, as well as some promising future extensions.

6.1 Possible Use Scenarios

Transferring to Other NLP Tasks Although we mainly demonstrated its usefulness with question-answering tasks, the nature of compressing documents into rectangle rows in flow maps can be expanded to many extensions beyond the task of question-answering. It may benefit many NLP tasks involving long sentences/phrases, especially those unbalanced ones. For example, in a summarization context, the visual system can show the alignment across different locations within the text to be summarized and the summarized text. At the same time, it can demonstrate the characteristics of each word token in the original document that helps readers understand how well their attention model performs.

As a Supporting Component In isolation, this individual visualization provides initial insights of how a relatively deep model handles instances. Moving beyond, this visualization may well serve as a plug-in for a system of larger scale, where it guides the generalization leading from specific examples to a broader overview.

For example, given a question starting with *when*, assuming the correct answer is yesterday night, but all attention were casted to type of tokens like a date. A reasonable hypothesis is then we overfit to *when* questions. These queries may serve as a start to identify data slices with questions starting with *when* while the corresponding answers are not in a specific date format. It is then very helpful to examine the model performance structurally so that we are able to deduce some actionable directions.

6.2 Limitations and Future Work

Further enhance the scalability. Although our presented visualization technique has addressed the issue of scalability in common question-answering tasks, squeezing even longer documents into one line of color rectangles could make each rectangle too small to interact with. Therefore, an important next step is to further declutter the flow map view. Intuitively, we can hierarchically group tokens, and collapse the less-relevant redundancy to make the visualization outcome more accessible. A more sophisticated way might be to take advantage of sequential matches (e.g., multi-token named entity phrases) so that the abstraction of individual token attentions into entropy-aware short phrases. Based on the entropies, we can provide overviews by encoding the attention distribution in each document as well as the mapping between them (e.g., one-to-one mapping, one-to-many mapping, or just random attention). These promising future directions could further improve the performance of our visualization model in its appropriate applications.

Support between-instance comparisons. Currently, our method focuses on visualizing one individual instance. We consider another critical future work on this system to be incorporating features to make the system more dynamic to situations when there were some perturbations on one of the multiple text sources. For example, in the context of question answering, various prior work has tried to analyze the through question-only or paragraph-only perturbations [15, 21]. Highlighting how the attention change with respect to the perturbation would strengthen our understandings on the model stability. As a starting point, such changes can be reflected on the heatmap: Instead of visualizing the absolute attention values, we can instead compute the relative changes, and distinguish those tokens getting more (less) attentions through color encodings.

Acknowledgments

We appreciate the valuable input from our study participants, and the constructive comments from the anonymous reviewers.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*, 2014.

[2] K. Buchin, B. Speckmann, and K. Verbeek. Flow map layout via spiral trees. *IEEE transactions on visualization and computer graphics*, 17(12):2536–2544, 2011.

[3] F. Dalpiaz, I. Van der Schalk, and G. Lucassen. Pinpointing ambiguity and incompleteness in requirements engineering via information visualization and nlp. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*, pp. 119–135. Springer, 2018.

[4] D. Guo and X. Zhu. Origin-destination flow data smoothing and mapping. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2043–2052, 2014.

[5] E. Hoque and G. Carenini. Convis: A visual text analytic system for exploring blog conversations. In *Computer Graphics Forum*, vol. 33, pp. 221–230. Wiley Online Library, 2014.

[6] M. Krauthammer and G. Hripcsak. A knowledge model for the interpretation and visualization of nlp-parsed discharged summaries. In *Proceedings of the AMIA Symposium*, p. 339. American Medical Informatics Association, 2001.

[7] J. Li, X. Chen, E. Hovy, and D. Jurafsky. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*, 2015.

[8] J. Li, W. Monroe, and D. Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*, 2016.

[9] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[10] Y. Ming, S. Cao, R. Zhang, Z. Li, Y. Chen, Y. Song, and H. Qu. Understanding hidden memories of recurrent neural networks. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 13–24. IEEE, 2017.

[11] C. Olah and S. Carter. Attention and augmented recurrent neural networks. *Distill*, 2016. doi: 10.23915/distill.00001

[12] D. Phan, L. Xiao, R. Yeh, and P. Hanrahan. Flow map layout. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 219–224. IEEE, 2005.

[13] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv:1606.05250*, 2016.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin. ”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, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144, 2016.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin. Semantically equivalent adversarial rules for debugging nlp models. In *Association for Computational Linguistics (ACL)*, 2018.

[16] M. Rikters, M. Fishel, and O. Bojar. Visualizing neural machine translation attention and confidence. *The Prague Bulletin of Mathematical Linguistics*, 109(1):39–50, 2017.

[17] A. Rücklé and I. Gurevych. End-to-end non-factoid question answering with an interactive visualization of neural attention weights. In *Proceedings of ACL 2017, System Demonstrations*, pp. 19–24, 2017.

[18] A. See, P. J. Liu, and C. D. Manning. Get to the point: Summarization with pointer-generator networks. *arXiv:1704.04368*, 2017.

[19] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016.

[20] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics*, 24(1):667–676, 2017.

[21] T. Wu, M. T. Ribeiro, J. Heer, and D. S. Weld. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 747–763, 2019.