Automatic Contextual Storytelling in a Natural Language Corpus

Ishrat Rahman Sami
isami001@gold.ac.uk
Supervised by Dr Tony Russell-Rose and Dr Larisa Soldatova
Goldsmiths, University of London
London, UK

ABSTRACT

Storytelling is an ancient art and science of conveying wisdom through generations for centuries. Data-driven storytelling in the context of a natural language corpus has a huge potential for conveying fast valuable insights about the corpus for better decision making. But high dimensional unstructured nature of natural language text makes automatic extraction of stories extremely difficult. This PhD research project believes that modern storytelling is a hand in hand approach of contextual topic visualization and contextual summarization. While exploratory data visualization can provide valuable insights into the data, these insights can be used to understand and design models for producing abstract summarization. In this project, the context of a story is defined from three perspectives: a single document, a collection of multiple documents about a topic of interest and the whole corpus. In this project, exploratory data visualization is used to understand the context better and now with the achieved insights, research is focusing on abstract summarization for automatic contextual storytelling.

CCS CONCEPTS

• **Human-centered computing** → **Visualization**; • **Computing methodologies** → **Natural language processing**; *Information extraction*; *Topic modeling*; • **Information systems** → *Summarization.*

KEYWORDS

Topic visualization; Abstract summarization; Storytelling; Natural Language Processing

ACM Reference Format:
Ishrat Rahman Sami. 2020. Automatic Contextual Storytelling in a Natural Language Corpus. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3418507

1 INTRODUCTION

Stories are a powerful medium for conveying wisdom. According to philosopher Maurice Merleau-Ponty, stories reveal information that we know about but we didn't know we knew [16]. Storytelling conveys an experience in a realistic manner. Storytelling can be used as a tool for gathering wisdom from a corpus, especially regarding knowledge, attitude and behaviour [14]. Storytelling with data representation differs from traditional classical approaches due to the complexity of the underlying content that needs to be communicated [20].

Contextual visualization and summarization have massive potential for revealing stories hidden in unstructured text. Journalists are increasingly using integrated data visualization in the form of diagrams and charts for supporting their storytelling [20]. In current time data visualization can be engaging and interactive by providing guided virtual tours through the analysis space [20]. Natural language text data is a raw unit of text mining. Wisdom about a corpus is more than presenting and arranging data, it is the applied knowledge about where and how to use the data [7]. Curating wisdom from raw data is a journey through acquiring factual information about the data and cultivating knowledge from the accumulated information [7]. Data visualization can play a massive role in identifying pearls of wisdom in a corpus. For better data visualization, compromises need to be made regarding dimensions. This limitation can be overcome by summarization. Document summarization can generate summary reflecting the most important points of a story which has been applied in biomedical, thread summarization, patent document analysis, etc. [26].

In this PhD project, context is defined from three perspectives:

- *A single document context* which is an ordered collection of sentences and represented by a set of words.
- *A corpus context* which is a collection of all documents in the corpus and represented by a set of weighted representative words or topics identified during single document analysis.
- *A relative context* which is a collection of multiple documents about a topic of interest and represented by a set of topics identified during single document analysis.

For contextual storytelling, this PhD project is exploring the opportunities for data visualization and document summarization.

2 PROBLEM

In this age of information flood, for informed directed decision making, it is vital for various professionals to understand the contextual affiliation of their topic of interest in their subjective corpus. But limited cognitive abilities of the human being make it impossible to deal with a massive amount of resources in a natural language corpus [6]. Automatic contextual storytelling can provide them with valuable insights. But the high dimensional nature of language makes contextual visualization and summarization extremely challenging. Some research regarding storytelling and summarization

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3418507

has been found [6][26]. Virtually all recent work on abstractive summarization considered single document [11]. Despite recent advances in summarization, the state of the art results for producing informative, fluent, and readable summary are not satisfactory[11]. Therefore, this PhD project is focusing on contributing to this research gap by answering the research question "What are the most effective data visualization and summarization techniques for increasing user understanding of key topics & trends in natural language texts?".

3 STATE OF THE ART

Data visualization and document summarization have important roles regarding contextual storytelling.

3.1 Data visualization

High dimensional natural language topics visualization usually use various scatter plots for cluster analysis. Principal Component Analysis (PCA) [17], t-distributed Stochastic Neighbor Embedding (t-SNE) [13] and variants of PCA and t-SNE are used to reduce dimension before producing scatter plots to analyze word embedding. Apart from dimension reduction techniques, sometimes researchers use topology driven approach to analyze hidden patterns in the data. Topology is a branch of mathematic that uses geometry and algebra for analyzing the structure of data [15][25]. Uniform Manifold Approximation and Projection (UMAP) [15] is a topology driven approach for dimension reduction. UMAP algorithm is a competitive visualization algorithm with t-SNE and has a better run time performance [25]. Apart from the word-embedding cluster visualization, a variety of chart visualization, typographic visualization and graph visualization are popularly used for their simplicity and intuitiveness [12]. Word cloud based steam graph can be used to demonstrate topic evolution [3]. ThemeRiver timeline can be used for displaying topic strength and bubble charts can be used to represent corpus documents [24]. Timelines have been historically used for storytelling regarding biographies, historical summaries, project plans etc. [1].

3.2 Document summarization

Document summarization can be classified as abstractive summarization and extractive summarization [6][26] and the task of summarization can be classified as generic summarization and query-oriented summarization [26]. This research is focusing on generic abstractive summarization which is a task of generating human alike summaries [8].

In early days various sentence ranking methods were used for extractive summarization followed by graph-based and template-based methods for abstract summarization [11]. With the significant development of neural network-based software and hardware technologies, the focus of text summarization has been shifted towards encoder-decoder sequence to sequence model based abstractive approaches in recent years [8][9][11]. Sequence to sequence models like Long Short Term Memory (LSTM) networks, Gated Recurrent Unit (GRU) networks [11], pointer generator networks [11], pointer generator with coverage are performing effectively in terms of ROUGE and BLEU evaluation matrices scores [8]. Rush et al. first incorporated the attention mechanism in 2015 for the sequence to sequence model and achieved the state of the art scores [18]. Attention-based mechanisms address the fact that some words/phrases are more important than others in the document [11]. The key objective of attention is to feed decoder an additional input called context vector that encodes important phrases [11]. Vaswani et al. introduced the transformer model in 2017 which replaces RNN with self-attention in encoder-decoder architecture and used multi-head attention to gain the state of the art performance [22]. This opened a new chapter of abstract summarization using transformers and its variants like two-staged transformer [21], RC-Transformer [2], transformer model with aggregation mechanism [9], etc.

Some research has been found regarding storytelling with summarization. Zhang, Ge and He used summarization for storytelling [26]. Their framework had three phases: document modelling using a weighted graph where sentences are vertices, sentence clustering for identifying latent topics in the story and sentence ranking for identifying story lines. Their experimental extractive results demonstrated the effectiveness of their approach. Janaszkiewicz et al. used sentence tagging based on sentiment classification for their storytelling exercise [6].

Large input size, re-usability and evaluation quality are the major problems for abstract summarization. Another problem of encoder-decoder architecture is that maximizing loss is not same as optimizing evaluation matrices[11]. Recently some researchers are using reinforcement learning for abstract summarization where an agent is trained to maximize reward based on evaluation metrics along with word information [11].

The results of abstract summarization can be evaluated via human review and/or automatic evaluation matrices. Manual human-based evaluation is time-consuming and expensive. Therefore, automatic evaluation methods like ROUGE[10] and its' variants ROUGE-N, ROUGE-L, ROUGE-SU etc. are mainly used for evaluating performance [11].

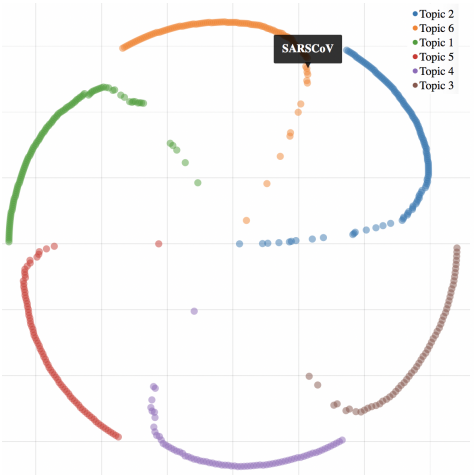

Figure 1: Peripheral contextual visualization of the corpus context for the article dataset COVID-19 bioRxiv/medRxiv [23]. It shows the 6 major trends and uses an interactive visualization. Hovering over the point exposes the topic.

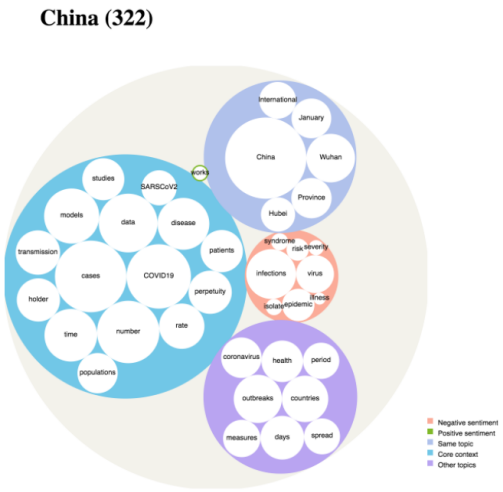

Figure 2: Relative contextualization of the word *China* considering 322 documents having topic *China* in the corpus COVID-19 bioRxiv/medRxiv corpus [23]. Bubble size represents the topics' weights. It gives us insights about the story of China in the corpus from the following perspectives:
- *How China is related to the COVID-19 corpus?* "Core context" (topics in a relative context that have more weighs in the corpus context than China)
- *Where/when in China COVID-19 happened?* "Same topic" (topics in a relative context that have the same LDA topic group)
- *Why we should care about China?* "Other topics" (topics in a relative context that have different LDA topic group)
- *What are the risk factors?* "Negative sentiment" (topics that belong to negative sentiment group)
- *What are the contributing factors?* "Positive sentiment" (topics that belong to positive sentiment group)

4 APPROACH

This research approached the automatic storytelling challenge from two perspectives:

- *Data visualization:* It is used as an exploratory tool for understanding hidden pattern of the story context.
- *Document summarization:* Based on the visualization insights, we are aiming to isolate language training from contextual topic discovery by masking the representative topics in source and target sequences.

For contextual analysis, primarily news/articles/book corpora have been chosen. Documents in these corpora follow a specific **writing pattern** which if visualized peripherally provide analytic insights [19]. Various studies on readers show that they usually skim the content by graphics, headline and initial paragraphs before starting reading the content [4] [5]. This **natural engagement of a reader** with news or articles motivates authors to provide worthy materials at the beginning which motivates readers to stop scanning and start reading [4]. During the data visualization stage, we attempted to capture this writing pattern peripherally based on position and occurrence of the word for a single document context. This representation was particularly useful for identifying topics using peripheral filtering. This peripheral topology is also persistent in corpus context which is useful for monitoring topic evolution. These findings were introduced in the paper 'A simplified topological representation of text for local and global context' [19].

To improve contextual trend analysis, we incorporated Latent Dirichlet Allocation (LDA) to group topics identified during single document analysis. Representing corpus topics by groups in a peripheral manner (Figure 1), not only preserves topic evolution symptoms but also reveals the trend of the corpus where time-series information is irrelevant. Proximity of the topics along the radius displays relatedness among the topics and proximity related to core shows the importance of the topic in the topic trend. Central topics are more researched topics than peripheral topics. Topics are grouped by LDA and the group number is set based on user input. This form of representation can be particularly useful for identifying the knowledge gap if intended to be used as an indicative tool for comparing trends against goals of the corpus. If topics related to goals are missing in the core, then more research work is related in the topic area.

Peripheral pattern is also consistent in a relative context for a topic of interest. Combining the information found in relative context and corpus context revealed insights of the story of a topic considering a point of view and a question in the human mind. The revealed insights are different from Open Knowledge Maps' thematic navigation system. We are attempting to publish a paper explaining these findings. The gathered insights from the visualization trigger the fact that if contextual topics can be masked from the summary and training a transformer for only learning how to write a summary gains a state of art performance then building reusable corpus independent summary generators will become easier. The process of combining behavioural pattern recognition using exploratory visualization and using the identified pattern in masked abstract summarization approach in the considered domains makes this approach and the findings unique and novel.

5 METHODOLOGY

The experimental methodology has been used for this research. Exploratory visual analysis of the defined contexts has been done during visualization. As the approach of contextual peripheral topological visualization is unique, this research lacks direct comparison against existing approaches. Attempts to compare peripheral visualization against word-embedding based t-SNE was proved to be unfair. Therefore, the evaluation of the visualization task is performed by humans. This research is currently focusing on the abstract summarization based on vanilla transformer model. During the writing of the paper, we are trying to isolate contextual topics from generic summary before and after training by masking and unmasking. We are aiming to use ROUGE for automatic evaluation of the generated summary.

6 RESULTS

The achievements of this research are as follows:

- We identified a consistent peripheral topological pattern in all the defined contexts.
- Identified peripheral topology can be used to filter topics in a single document context [19].
- Peripheral topology in corpus context can be used for monitoring topic evolution [19]. It can also be useful for identifying trends and knowledge gap in the corpus. Figure 1 presents an example.
- Peripheral topological pattern is also consistent in a relative context. Comparing relative context against the corpus context reveals the elements of the story of the topic in the corpus as explained in Figure 2 for the topic 'China' in the corpus COVID-19 bioRxiv/medRxiv [23].

According to Janaszkiewicz et al., the 7 most important elements for digital storytelling are the point of view, dramatic question, emotional content, the power of soundtrack, economy (using only enough information), pacing the rhythm of the story and a storyteller [6]. Looking at Figure 2 from the point of view of a researcher, with the question, "What is China's role in COVID-19 article knowledge base?" in mind, the representation indeed provides insights about COVID-19 outbreak story of China economically. "Core context" in Figure 2 provides the insight that "China is related to the knowledge base because of COVID-19 disease which is also called SARSCoV2. China's contribution to the knowledge base is via the studies and data collected from its infected population during transmission." "Same topic" in Figure 2 provides the insight that "There was an international outbreak in January in Wuhan, Hubei Province, China." "Other topic" in Figure 2 provides the insight that "Researchers should be concerned about coronavirus related health outbreaks and measures need to be taken to stop the spread." "Negative sentiment" in Figure 2 provides the risks insights got from China which are infections, virus, epidemic, severity, isolation, etc. "Positive sentiment" in Figure 2 provides the fact that the research works done during China's outbreak was vitally important initially. We got similar results about other topics in the corpus.

The main challenge of the visualization task was user evaluation. Due to the uniqueness of this approach, we haven't managed to directly compare our outputs against existing approaches. Integrating user survey might be useful to prove the richness of these visualizations.

7 CONCLUSION AND FUTURE WORK

Language is a powerful tool for communication which is not limited to words, semantics, sentiments and classification of topics. Language conveys personality of author/speaker and establishes a relation between a person and relative audience. This communication behaviour influences writing patterns and exploratory visualization was extremely useful for identifying the value and usefulness of peripheral writing patterns in the specified contexts. Stories bring understanding of a context by providing a motive through a point of view. Stories connect the topics in a fluent manner. This research is uniquely attempting to isolate language training from contextual understanding for fluent abstract storytelling.

REFERENCES

[1] Matthew Brehmer, Bongshin Lee, Benjamin Bach, Nathalie Henry Riche, and Tamara Munzner. 2017. Timelines Revisited: A Design Space and Considerations for Expressive Storytelling. *IEEE Transactions on Visualization and Computer Graphics* 23, 9 (2017), 2151–2164.

[2] Tian Cai, Mengjun Shen, Huailiang Peng, Lei Jiang, and Qiong Dai. 2019. Improving Transformer with Sequential Context Representations for Abstractive Text Summarization. In *CCF International Conference on Natural Language Processing and Chinese Computing*. 512–524.

[3] Tommy Dang, Huyen N Nguyen, Vung Pham, J Johansson, F Sadlo, and GE Marai. 2019. WordStream: Interactive Visualization for Topic Evolution. In *EuroVis*.

[4] Mario R Garcia, Pegie Stark, and Ed Miller. 1991. *Eyes on the News*. Poynter Institute for Media Studies St. Petersburg, FL.

[5] Kenneth Holmqvist, Jana Holsanova, Mari Barthelson, and Daniel Lundqvist. 2003. Reading or scanning? A study of newspaper and net paper reading. In *The Mind's Eye*. Elsevier, 657–670.

[6] Piotr Janaszkiewicz, Justyna Krysińska, Marcin Prys, Magdalena Kieruzel, Tomasz Lipczyński, and Przemysław Różewski. 2018. Text Summarization For Storytelling: Formal Document Case. *Procedia Computer Science* 126 (2018), 1154–1161.

[7] Ali Fenwick Jose Berengueres, Marybeth Sandell. 2019. *Introduction to Data Visualization & Storytelling A Guide For The Data Scientist*. ResearchGate.

[8] Heena Kumari, Sunita Sarkar, Vikrant Rajput, and Arindam Roy. 2020. Comparative Analysis of Neural Models for Abstractive Text Summarization. In *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences*. 357–368.

[9] Pengcheng Liao, Chuang Zhang, Xiaojun Chen, and Xiaofei Zhou. 2019. Improving Abstractive Text Summarization with History Aggregation. *arXiv preprint arXiv:1912.11046* (2019).

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

[11] Hui Lin and Vincent Ng. 2019. Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9815–9822.

[12] Shixia Liu, Xiting Wang, Christopher Collins, Wenwen Dou, Fangxin Ouyang, Mennatallah El-Assady, Liu Jiang, and Daniel A Keim. 2018. Bridging text visualization and mining: A task-driven survey. *IEEE transactions on visualization and computer graphics* 25, 7 (2018), 2482–2504.

[13] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[14] Becky McCall, Laura Shallcross, Michael Wilson, Christopher Fuller, and Andrew Hayward. 2019. Storytelling as a research tool and intervention around public health perceptions and behaviour: a protocol for a systematic narrative review. *BMJ open* 9, 12 (2019).

[15] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[16] Maurice Merleau-Ponty et al. 1964. *The primacy of perception: And other essays on phenomenological psychology, the philosophy of art, history, and politics*. Northwestern University Press.

[17] Bruce Moore. 1981. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE transactions on automatic control* 26, 1 (1981), 17–32.

[18] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685* (2015).

[19] Ishrat Rahman Sami and Katayoun Farrahi. 2017. A simplified topological representation of text for local and global context. In *Proceedings of the 25th ACM international conference on Multimedia*. 1451–1456.

[20] E Segel and J Heer. 2010. Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 1139–1148.

[21] M. Su, C. Wu, and H. Cheng. 2020. A Two-Stage Transformer-Based Approach for Variable-Length Abstractive Summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2020), 1–1.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[23] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. CORD-19: The Covid-19 Open Research Dataset. *arXiv preprint arXiv:2004.10706* (2020).

[24] Yi Yang, Quanming Yao, and Huamin Qu. 2017. VISTopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. *Visual Informatics* 1, 1 (2017), 40–47.

[25] Julio Christian Young and Andre Rusli. 2019. Review and Visualization of Facebook's FastText Pretrained Word Vector Model. In *2019 International Conference on Engineering, Science, and Industrial Applications (ICESI)*.

[26] Zhengchen Zhang, Shuzhi Sam Ge, and Hongsheng He. 2012. Mutual-reinforcement document summarization using embedded graph based sentence clustering for storytelling. *Information Processing and Management* 48, 4 (2012), 767–778.