An Empirical Study on Topic Preservation in Multi-Document Summarization

Mong Yuan Sim and **Wei Emma Zhang** and **Congbo Ma**
University of Adelaide
Adelaide SA 5005
mongyuan.sim@student.adelaide.edu.au,
{wei.e.zhang, congbo.ma}@adelaide.edu.au

Abstract

Multi-document summarization (MDS) is a process of generating an informative and concise summary from multiple topic-related documents. Many studies have analyzed the quality of MDS dataset or models, however no work has been done from the perspective of topic preservation. In this work, we fill the gap by performing an empirical analysis on two MDS datasets and study topic preservation on generated summaries from 8 MDS models. Our key findings include i) Multi-News dataset has better gold summaries compared to Multi-XScience in terms of its topic distribution consistency and ii) Extractive approaches perform better than abstractive approaches in preserving topic information from source documents. We hope our findings could help develop a summarization model that can generate topic-focused summary and also give inspiration to researchers in creating dataset for such challenging task.

1 Introduction

Multi-document summarization (MDS) is a task to produce an informative and concise summary from multiple documents. In general, there are two different approaches to MDS, which are extractive and abstractive summarization (Ma et al., 2022). Extractive summarization refers to methods that select important sentences from input documents and produce a summary. These methods perform better at producing summary without grammatical errors. On the other hand, abstractive summarization refers to methods that have the ability to generate summaries with words that do not exist in input documents (Cui and Hu, 2021; Fabbri et al., 2019).

The development of text summarization model has been supported by the growing amount and quality of available dataset. The available dataset types vary from news articles (Fabbri et al., 2019) to scientific articles (Lu et al., 2020) and Wikipedia abstract (Perez-Beltrachini et al., 2019). However, information is not scarce in this era, but "valuable" information is. Many recent work (Cui and Hu, 2021; Zou et al., 2021; Zhu et al., 2021; Perez-Beltrachini et al., 2019) have been focusing on generating topic-guided summaries using one-size-fits-all dataset which are not meant for this kind of work, making it difficult to evaluate whether the model is performing better than "generic" model in terms of the quality of generated topic-focused summary. There are also work (Zhang et al., 2021; Tejaswin et al., 2021; Xu et al., 2020) focusing on the analysis of summarization models and datasets but none on topic-preservation. To the best of our knowledge, there is only one dataset (Bahrainian et al., 2022) created for topic-guided news summarization, but has been tailored for single document summarization. Therefore, it is essential to deep dive into current available MDS dataset, and investigate their suitability for developing topic-guided summarization models, and the pattern of high quality summaries in order to inspire future work in text summarization.

In this paper, we conducted several experiments in analyzing the relevance of input and output documents in automated summarization and the pattern of model-generated summaries. To sum up, our contributions are two-folds: i) for MDS dataset, we evaluated topic relation between source documents and gold summaries in widely used MDS dataset, inspiring future work on creating high quality dataset for topic-aware summarization model; ii) for MDS models, we investigated summaries generated from a wide range of state-of-the-art models in order to provide insights of how relevant it is to the source documents. Our observations could inspire research directions towards better topic-preserving MDS dataset and models.

2 Datasets and Models

In this work, we use two most commonly used multi-document summarization dataset Multi-

Proceedings of the AACL-IJCNLP 2022 Student Research Workshop, pages 61–67
November 20, 2022. ©2022 Association for Computational Linguistics

News and Multi-XScience in our experiments. We run 8 MDS models from non-deep learning based models to deep learning models including the recent Transformer-based state–of-the-art models. For fair comparison, training/validation and testing for all models are performed on a high performance computing cluster powered by NVIDIA V100.

2.1 MDS Datasets

2.1.1 Multi-News

Multi-News (Fabbri et al., 2019) is the first large-scale dataset constructed by collecting human-written articles which are summaries of multiple news article sources from newser.com. This dataset contains 44,972/5,622/5,622 instances for training, validation, and testing. Each instance has 2 to 10 source documents per summary.

Source documents and gold summaries for Multi-News are stored in different .txt files. In source documents file, documents used to generate one summary are separated by a token called "story_special_token_tag". We processed the dataset by removing the token to separate source document and unused words such as "<unk>" and "<blank>" before feeding them into topic model.

2.1.2 Multi-XScience

Multi-XScience (Lu et al., 2020) is a large-scale dataset created for extreme summarization task which is to write related-work section of a paper based on its abstract and the articles it references. Information is collected from arvix.org and Microsoft Academic Graph (MAG). This dataset contains 30,369/5,066/5,093 instances for training validation, and testing. Each instance has 10 to 20 references as input.

Multi-XScience dataset comes in as a JSON file. Each data instance contains a related work section which is the gold summary, along with multiple "ref_abstract" entries which act as source documents. The citation in the sources and targets are replaced by a common token "@cite". We process the dataset by storing them in a list, remove unused words and tokens such as "@cite".

2.2 MDS Models

In order to examine model generated summary, we generate summaries from 8 MDS models including both extractive and abstractive models. The overview of these models are as follows:

MMR (Goldstein and Carbonell, 1998) is an extractive approach that assigns scores to sentences and re-rank them to obtain relevant sentences.

Textrank (Mihalcea and Tarau, 2004) produces undirected weighted graph from input documents, focusing on keywords to find the most relevant sentences in text.

Lexrank (Erkan and Radev, 2004) is an extractive method that uses graph-based method to compute relative importance of documents.

PG (See et al., 2017) pointer-generator model extends the standard seq2seq framework with copy and coverage mechanism.

Transformer (Vaswani et al., 2017) captures cross-document relationships via attention mechanism.

CopyTransformer (Gehrmann et al., 2018) randomly chooses one of the attention heads of Transformer as the copy distribution.

Hi-Map (Fabbri et al., 2019) adapts a pointer-generator model with MMR to compute weights over multiple documents inputs.

SummPip (Zhao et al., 2020) converts documents to sentence graph, apply spectral clustering to obtain clusters of sentences.

3 Methods and Results

We compare and analyze topic-related patterns of source document, gold summaries (provided in MDS benchmark datasets), and the generated summaries (from 8 MDS models). Guided by topic modelling research, we adopt the topic related evaluation metrics in this work. We specifically study i) topic coherence, to identify the best settings; ii) number of documents in each topic, to study the overall topic distribution; iii) distances among topic distributions of summaries, to examine the document-level patterns; iv) topic words correlations in summaries, to analyze the word-level patterns.

3.1 Latent Dirichlet allocation (LDA)

Latent Dirichlet allocation (LDA) is a generative probabilistic topic model for which each document is represented as a random mixture of latent topics and each topic is represented as a distribution over fixed set of words (Onan et al., 2016). It aims to identify underlying latent topic structure based on observed data (Blei et al., 2003).

We make good use of information obtained from a LDA based topic model, which are topic distribution and word vector for each topic. Topic distribution is a vector contains N elements, where

N is number of topics. Each value represents the probability of a document falls into topic group n. Word vector shows the weight for each word in a topic:

$$\begin{aligned} doc_i &: [w_{tp1}, w_{tp2}, ..., w_{tpN}] \\ tp_j &: [w_{wd1}, w_{wd2}, ..., w_{wdM}], j \in [1, N] \end{aligned} \quad (1)$$

where w_{tpj} is the probability of the j-th topic in the i-th document, w_{wdq} is the probability of q-th word in j-th topic, M is the number of words to describe a topic. Both N and M are hyper-parameters.

We apply LDA topic modelling on the corpus containing all the source documents, gold summaries and generated summaries as they are in the same topic distribution space and we want to observe the topic patterns within.

3.2 Topic Coherence

Topic coherence is a qualitative measurement to measure the quality of topic modelling (Newman et al., 2010). The underlying idea is rooted in the distributional hypothesis of linguistics that consider words with similar meanings tend to occur in the similar contexts (Harris, 1954). If a topic's top K words have related meanings, the topic is considered to be coherent (Syed and Spruit, 2017).

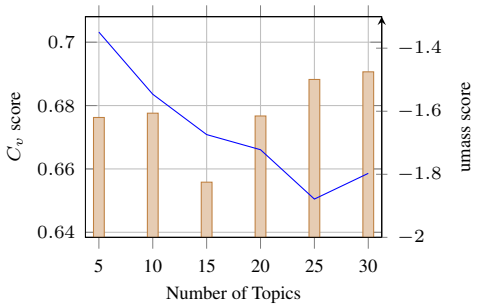
In this study, we use topic coherence score to identify the best hyper-parameter settings for the topic model LDA, and use this setting for follow-up experiments. Particularly, we adopt the coherence measure proposed by Röder et al. (2015) (known as UMass-coherence) which is calculated based on co-occurrences of word pairs as follows:

$$C_{UMass}(T) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{p(w_m, w_l) + \frac{1}{|D|}}{p(w_l)} \quad (2)$$

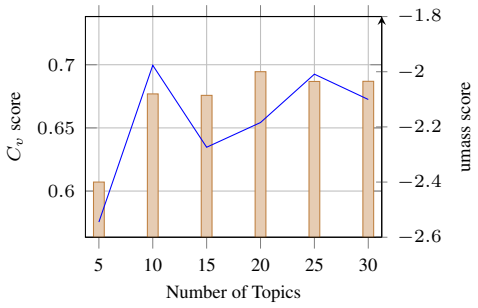
where $p(w_m, w_l)$ denotes the probability of the co-occurrence of words w_m and w_l in the corpus D. It is computed as the ratio of number of documents containing both words w_m and w_l and the total number of documents in D. M is the length of the word list.

Another commonly used topic coherence score is C_v score, which creates content vectors of words using word co-occurences and calculates the score using normalized pointwise mutual information (PMI) and cosine similarity.

We obtain topic distribution and word vector for each topic on both dataset, Multi-News and Multi-XScience from LDA. Then we identify the best topic coherence score for these two datasets in order to get the best topical setting. We compute the two types of coherence scores, namely u_mass and C_v score as discussed previously. The number of topics is set to 5, 10, 15, 20, 25 and 30. From the results shown in Figure 1, we observe that when the number of topics is 25, Multi-News dataset shows the highest coherence score. For Multi-XScience dataset, 5 topics achieves best coherence. We use these settings for the follow-up experiments.



(a) Multi-News



(b) Multi-XScience

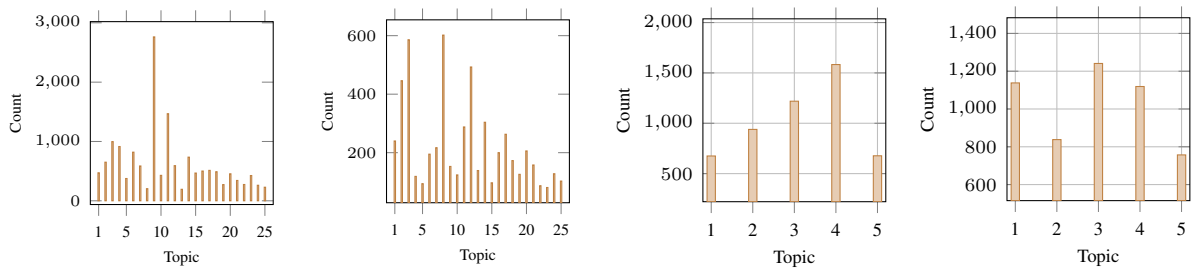
Figure 1: Topic coherence score for Multi-News and Multi-XScience datasets. The bar chart represents C_v score while line chart represents umass score.

3.3 Analyzing number of documents per topic

To discover the overall topic distribution of the dataset, we perform K-Means clustering on the topic distribution obtained from LDA model. We notice that in Figure 2a and 2b, Multi-News source documents are "heavy" in topic 9 while its gold documents mostly fall into topic 1, 2, 3, 8 and 12. For Multi-XScience, although it does not show domination by any topic, we can still see from Figure 2c and 2d that source documents and gold summaries do not follow the same topic distribution.

3.4 Distances of Topic Distributions

We measure the distances of document-topic distributions of source documents, gold summaries and the generated summaries, aiming to find document-level topical correlations.



(a) Multi-News (Source) (b) Multi-News (Gold) (c) Multi-XScience (Source) (d) Multi-XScience (Gold)

Figure 2: Number of Documents per Topic for Multi-News and Multi-XScience.

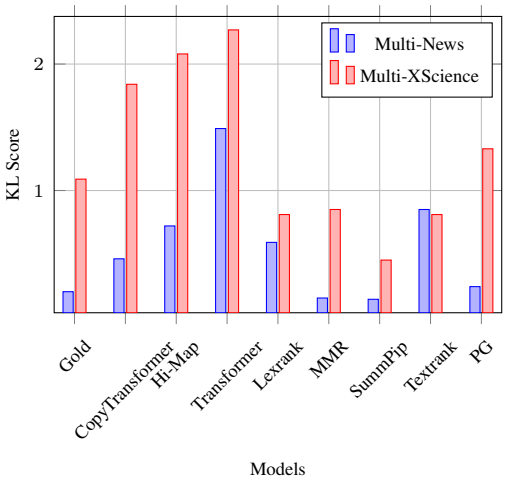


Figure 3: KL score for MultiNews and Multi-XScience. The first two bars shows the KL score between source documents and gold summaries. The rest of the bars show KL score between source documents and generated summaries as labeled.

We adopt Kullback-Leibler (KL) Divergence measure as the distance function. KL-divergence is a way to quantify the distances between two probability distributions (Shlens, 2014). Given two probability distribution density functions (PDFs), p and q, their KL divergence score, denoted as $KL(p \parallel q)$, is defined as :

$$KL(p \parallel q) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} dx \quad (3)$$

As we are focusing on multi-document summarization, the relationship between source document and summary is many to one. We first calculate the average topic distribution over all source documents, then compute KL divergence between source document and summary.

We present the document-level distances between source documents and summaries (gold and generated) in Figure 3. We can see that extractive models such as Lexrank, MMR, SummPip and Textrank tend to produce a summary where its topic distribution is closer to the source documents. On the other hand, the abstractive models which have proven to achieve higher ROUGE score, failed to produce a summary that is topic-relevant to the source documents. Transformer, one of the most popular trends in summarization has the highest KL score on both dataset which means summaries produced from this model are often "off-topic" in a sense that it fails to capture the underlying topic.

We also observe that Multi-News dataset provides gold summary that preserves topic information better than Multi-XScience. Overall, Multi-News has lower KL score in both gold and generated summaries compared to Multi-XScience.

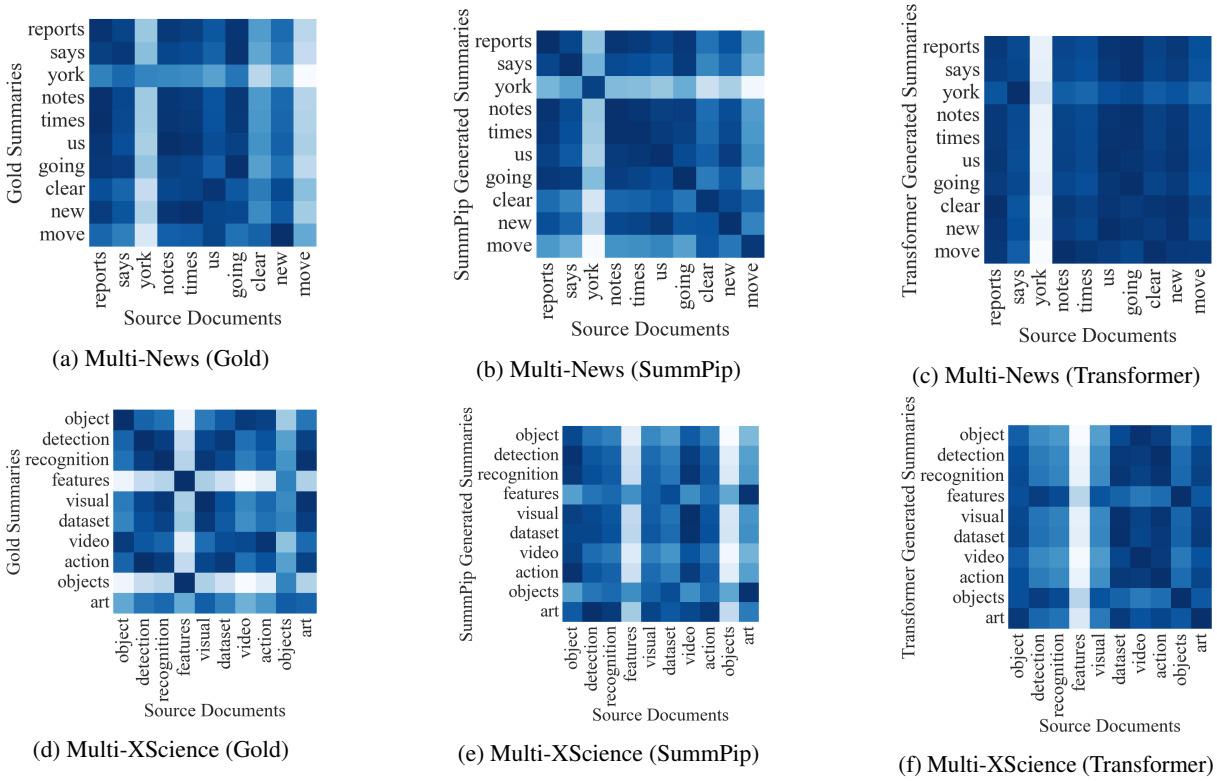
3.5 Topic Words Correlation

As we want to explore the correlation between source documents and summaries, along with gold and model generated summaries, we compute a new weight of each word in a document by multiplying topic weight by word matrix for each topic. The resulting vector shows the weight for each word in a document. For example, the q-th word in the j-th topic of document i has weight $w_{tpj} * w_{wdq}$. Then we consider the correlation of words in two document as the euclidean distance of their weights.

We obtain the words' weights in a document by using their topical probabilities. We depict the word correlations between source documents, gold summaries and summaries generated from SummPip and Transformer in a heatmap. We selected SummPip and Transformer because they have the lowest and highest KL score respectively.

For visualisation purpose, we picked top-10 words from a topic and computed euclidean distance of each word in two vectors. If two documents are highly correlated,the heatmap will have a straight line from top left to bottom right.

From Figure 4, we can see that the result we obtain is very far away from best case scenario.



(a) Multi-News (Gold)

(b) Multi-News (SummPip)

(c) Multi-News (Transformer)

(d) Multi-XScience (Gold)

(e) Multi-XScience (SummPip)

(f) Multi-XScience (Transformer)

Figure 4: Word-level correlation for source documents and summaries for Multi-News and Multi-XScience dataset. We randomly select a topic and visualise the correlation between source documents and summaries with top-10 words for that topic. The higher the correlation, the darker the square is.

Instead, the words sparse across many different topics and are inconsistent. This means a model generated summary might be discussing about a very different topic than those in source documents.

4 Discussions

Our studies and observations raise the following questions that we believe need to be considered in the MDS research:

Is extractive MDS model better than abstractive MDS model in preserving topics? From results in Section 3.4, we find that in terms of topic preservation, extractive models work better than abstractive models. This could due to the "extract" nature of the former which shares the same vocabulary as the source document, resulting in higher word correlation between source documents and summaries. Future work could focus on analyzing word semantic similarities instead of relying on topic distribution similarities only as abstractive models use words that are different from source documents to generate a summary. To improve the topic preservation of abstractive models, we could consider select semantically similar words to the words in the source document during generation.

Whether gold summary follows source documents' topic distribution? From Section 3.4 we also find Multi-News's gold summaries topic distribution are well aligned with the topic distributions in its source document, however Multi-XScience does not perform well in this regard.This analysis could inspire future MDS dataset contributors to take topic preservation into consideration when preparing gold summaries such that source documents and gold summaries have similar topic distribution.

Whether the number of documents are similar across all topics? Dataset that is "heavy" on one topic can disadvantage summarization models in training as the vocabulary might be dominated by a specific topic causing topic information for other topics with less instances being discarded or normalised. This can be seen in Figure 2 where the document count per topic for source documents and gold summaries are inconsistent. Future dataset creation should focus on the topic distribution among all documents in data collected to make sure that the generation model captures equal information from all topics.

5 Conclusion and Future Work

In conclusion, we have systemically and empirically analyzed two popular multi-document summarization datasets and summaries generated from a variety of state-of-the-art summarization models. Our analysis over 100,000 documents reveals that source documents, gold summaries and model generated summaries are rarely topic coherent which cause the summary to be less informative for some usages. This analysis also lead to some inspiration and suggestions in creating better summarization models and dataset for real world application.

References

Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. 2022. NEWTS: A corpus for news topic-focused summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 493–503, Dublin, Ireland. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Peng Cui and Le Hu. 2021. Topic-guided abstractive multi-document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1463–1472, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Jade Goldstein and Jaime G. Carbonell. 1998. Summarization: (1) using MMR for diversity- based reranking and (2) evaluating summaries. In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, MD, USA, October 13-15, 1998*, pages 181–195. Morgan Kaufmann.

Zellig S. Harris. 1954. Distributional structure. *WORD*, 10(2-3):146–162.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8068–8074. Association for Computational Linguistics.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2022. Multi-document Summarization via Deep Learning Techniques: A Survey. Accepted at March 2022.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 404–411. ACL.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, Los Angeles, California. Association for Computational Linguistics.

Aytug Onan, Serdar Korukoglu, and Hasan Bulut. 2016. Lda-based topic modelling in text sentiment classification: An empirical analysis. *Int. J. Comput. Linguistics Appl.*, 7(1):101–119.

Laura Perez-Beltrachini, Yang Liu, and Mirella Lapata. 2019. Generating Summaries with Topic Templates and Structured Convolutional Decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5107–5116, Florence, Italy. Association for Computational Linguistics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015*, pages 399–408. ACM.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Jonathon Shlens. 2014. Notes on Kullback-Leibler Divergence and Likelihood. ArXiv:1404.2000 [cs, math].

Shaheen Syed and Marco Spruit. 2017. Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 165–174, Tokyo, Japan. IEEE.

Priyam Tejaswin, Dhruv Naik, and Pengfei Liu. 2021. How well do you know your summarization datasets? In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3436–3449. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020. Understanding neural abstractive summarization models via uncertainty. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6275–6281. Association for Computational Linguistics.

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, and Dragomir R. Radev. 2021. An exploratory study on long dialogue summarization: What works and what's next. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4426–4433. Association for Computational Linguistics.

Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. Summpip: Unsupervised multi-document summarization with sentence graph compression. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1949–1952, New York, NY, USA. Association for Computing Machinery.

Fangwei Zhu, Shangqing Tu, Jiaxin Shi, Juanzi Li, Lei Hou, and Tong Cui. 2021. TWAG: A topic-guided Wikipedia abstract generator. 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)*, pages 4623–4635, Online. Association for Computational Linguistics.

Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14665–14673. AAAI Press.