ISSN 1533-9211

ABSTRACTIVE SUMMARIZATION USING CATEGORICAL GRAPH NETWORK

R. SENTHAMIZH SELVAN[1], Dr. K. ARUTCHELVAN[2]
[1]Research Scholar, [2]Assistant Professor,
[1,2]Department of Computer and Information Science, Annamalai University, Chidambaram, Tamil Nadu, India

DOI 10.5281/zenodo.6552514

Abstract:
The rapid development of technologies produce enormous amount of data which have lot of hidden insights. Extracting these hidden insights are challengeable for researchers and industrialists. Most of the data are in textual and unstructured format. Text mining is the prominent research area that has being utilized for the textual data analysis. Document summarization is an effective application which provides the summary of given content. This research work mainly focused on generating abstractive summarization from the multiple documents. It contributes abstractive summarization using the categorical graph network. Lot of duplicate or redundant sentences are there in the multiple documents. Proposed CATSum, which is a graph based abstractive summarization technique that identifies the duplication based on the similarities of the sentences. The proposed technique used ALBERT encoder model to train the datasets. Then it has built the content summary based on the connection between the sentences. The proposed work is measured using the ROUGE-1, ROUGE-2 and ROUGE-L metrics and produced better accuracy than the baseline methods.
Keywords: Abstractive Text Summarization, Categorical Graph Network, Multiple-document Summarization.

1. INTRODUCTION

Natural Language Processing (NLP) is widely being used in various kind of applications such as sentiment analysis, text summarization, chat-bot etc. Text summarization (TS) is the process that is used to provide brief content from a detailed contents [1]. TS systems can be categorized as single or multiple document,s summarization based on the number of input. The summary which can be extracted from a single source of document is referred as single-document summarization. Nowadays, lot of research works are ongoing for multi-document summarization for the similar topic [2], [3]. The summarization technique can be further classified into two main types that are as follows: Extractive Summarization and Abstractive Summarization [4]. Extracting important sentences and combining the important sentences that are extracted from the single/multi-document is known as extractive summarization [8]. Abstractive summarization uses antithetic words to ingeminate the meaning instead of choosing the actual sentence from raw information [5]. For a decade, many researchers have focused on extractive summarization (ES). It focuses on extracting the key information based on the sentence ranking. Query based summarization is being used nowadays. This summary

The Seybold REPORT

ISSN 1533-9211

the contents based on the query passed by the users [6], [7]. In general, Extractive Summarization provides a sentence that semantically and grammatically corrected [1] [2] and calculated fast.

Apart from this, the reinforcement learning has been established for considering the semantics of the separated summary or abstract [9] which integrates the maximum possible cross-entropy loss with the service provided by the rule slope to frankly improve the assessment metric for the abstract task. In recent, a popular solution has been created for abbreviation system with a two-level decoder. These solution extract key phrases, rewrite, abbreviate these sentences [10]. In general, preceding models utilize the top-k approach like the best approach to diverse documents, the amount of chosen sentences is fixed, which contradicts the actual world. For instance, approximately all preceding models take out three phrases from the actual articles [4] however, 40% of documents on CNN / DM are more or less than 3-sentences or abbreviations. This is because such models make it difficult to measure excitement and redundancy simultaneously with the spread of error.

At presents, most neural extractive modulation systems vote on the individual text one at a time from the real text and then model a number of sentences to create a synopsis. Narayan ET. Al [11], have created a extractive sequence labeling problem and solve it with an encryption decoder structure. These models achieve more redundancy because it makes autonomous dual conclusions for all sentences. To solve the above problem, auto-regressive decoder is used that permits the gaining functions of diverse sentences to control each other [10]. Moreover, the recent developments focus on balancing the specialization and redundancy of sentences, i.e. choose more semantic similar phrases to solve the encountered problems among chose phrases.

One of the most popular method which works on the similar motivation is Trigram blocking [7], At the point of selecting phrases to create the abstract, this will avoid the sentence where the trigram overlaps with the previously selected phrases. Amazingly, this method of duplicate removal brings significant enhancement in the performance to CNN /Daily Mail. The systems for developing the association among sentences are mostly sentence-level separators, regardless of the semantics of the whole summary. This gives them more choice to choose the most common phrases, while ignoring the link of multiple phrases. Reinforcement learning (RL) used to attain abstract level assessment, but is still limited to the structure of sentence-level summaries.

This research work focuses on the abstractive summarization for multi-documents with similar topic. The proposed CATSum (Categorical graph network based Summarization) technique uses categorical graph network for building the relationship among the sentences. Later the sentences are trained with the pre-trained ALBERT model. The proposed technique is evaluated using the most popular datasets that are available publicly. The detailed architecture of the proposed technique is discussed in section 2. Performance of the proposed technique is evaluated using the popular ROGUE metrics and it is presented in the section 3. Section 4 concludes this research work.

The Seybold Report
ISSN 1533-9211

2. CATSum TECHNIQUE

2.1 Introduction

The detailed overview of the research methods which are used for this research contribution is given in this section. The following sections explain the process of dataset collection and preparation methods. Technique proposed for this research is explain with the mathematical notations and algorithms in the consecutive section of this section.

Consider $Sen = \{sen_1, sen_2, \ldots sen_N\}$, which represents the collected multi-document's sequences that contains N number of sentences, where sen_i is i-th sentence of the collected document. Assume the human-generated summary as H. Abstractive summarization targets to yield the summaries $Sen^* = \{sen^*_1, sen^*_2, \ldots sen^*_N\}$ by choosing P sentences from Sen, where $P \leq N$. Labels $X = \{x_1, x_2, \ldots, x_N\}$ are resulting from H, where $x_i \in \{0,1\}$ represents whether sentence sen_i should be comprised in the extracted summary.

To demonstrate the repetition connection between the sentences, this research work utilizes a heterogeneous graph. As it is given in the figure 1 [9], the graph contains multi-granularity levels of data to represent the sentences. Three sorts of nodes such as NE (named entity), sentence and word are present in the proposed graph. The proposed technique replace the object that holds the information of the NE with the different tokens (e.g. [person_1], [person_2], [country_1], [org_1] and [date_1]) to reduce the semantic complexity. To represent the word-level data, word node is considered to process the graph. Most of the existing research work, eliminated the repeated words that contain the same meaning. In this proposed work, it keep the repeated or identical words as a separate node to avoid the interception between different circumstances. The relationship between the sentences is represents by the sentence node.

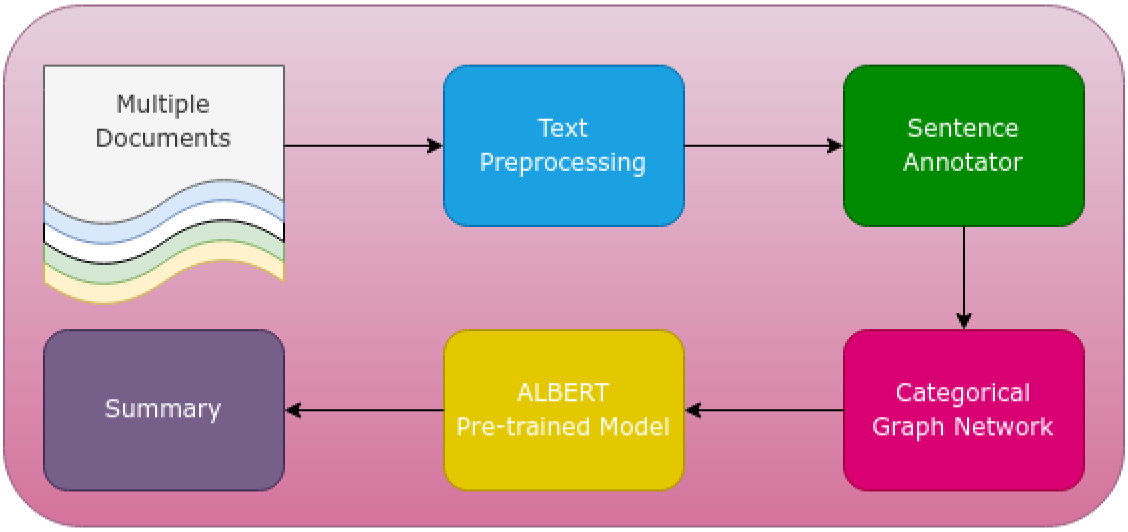

Figure 1 Architecture of CATSum

ISSN 1533-9211

Categorized Attentive Graph Network Summarization (CATSum) defines five types of edges to represent the different structural information that are as follows:

- Directed next edges: It is used to connect the sequential NE and words in a single sentence.
- Directed in edges: It is used to connect the NE node or word from one sentence node to another sentence node if the NE node is repeated in another sentence node
- Directed out edges: It is used to connect the sentence node that preceded from the previous sentence node using NE word appeared in different sentence node
- Undirected sentence_similar edge: It is used to connect the sentence nodes that has overlapping pharses (trigrams)
- Undirected NE_same edge: It is used to connect the NE if it is redundant.

Category levels of the graph are represented by adjacency matrix *Adj*, where the presence of edges among the nodes is indicated by Boolean value. The CATSum consists of four subgraphs that are as follows:

i. Word (A_{word}): Next edge and NE_same edge is used to construct graph between entity and word node.
ii. Word-named entity (A_{word_NE}): Directed in and directed out edge is used to build relationship between NE, word and sentence nodes.
iii. Word-sentence (A_{word_sen}): Sentence_similar and NE_same edges are used to build relationship between word and sentence nodes.
iv. Sentence (A_{sen}): This builds relationship between A_{word_sen} and sentence_similar edges. It is like nested edges.

2.2 Categorical Graph Network (CGN)

It is well known that each and every sentence in a document is interconnected with each other. The sentences are considered as nodes. Categorical graph network (CGN) is build to identify the importance of the sentences using node weight. The nodes in the graph are portrayed as an adjacency matrix. The matrix is denoted by $A_{adj} \in \mathbb{R}^{NXN}$where N denodes nodes. Each node 'n' is delineated as the categorical features. The categorical features are determined by using the named entity recognition. The CGN is performed by using the Unsmoothed Neighborhood Aggregation (UNA) method. The UNA identifies the proper neighborhood of each node. This helps to provide the content flow for generating the abstractive summary. The neighborhood aggregation of m-hop is given in the equation 1.

$$h_N^m(v) \leftarrow UNA_k\left(h_u^{m-1}, \forall u \in N(v)\right), (1)$$

Where v = vertex, u = node and N = number of nodes

In the neural network, softmax function plays an important role to sort the coefficients between the various nodes for an easy comparison. It is the final and output layer of neural network (NN) based classifier. The following equation (2) is the normalized of softmax:

The Seybold REPORT

ISSN 1533-9211

$$softmax(\vec{u}) = \frac{e^{u_i}}{\sum_{j=1}^{m} e^{u_j}} (2)$$

Where u_i represents the input vector and u_j refers output vector.

2.3 ALBERT Encoder

The proposed CATSum is based on the architecture BERTSUMEXT [12]. The graph is built with a variant of different edges. The ALBERT encoder [13] is used to learn the contextual representations of words. ALBERT is a pre-trained model that provides the summary using BERTSUMEXT model. For the hidden layer in the neural network, the ALBERT encoder considers the word and sentence nodes as the hidden layer. It takes subword as the input. In this research work average pooling function is used to produce the output from ALBERT encoder. The given inputs in the ALBERT encoder are analyzed by the following three layers:

- Abstract layer: It contains the GAN subgraphs. The word and sentence level subgraphs are processed and passed in to the redundancy layer.
- Redundancy layer: This layer works with the sentence level to identify the similar sentences and provides the scores.
- Output layer: The average pooling function is used to process the output from the above two layers.

The abstract layer is designed to learn the semantic recurrence of each word in the word-level diagram. Then, convert the word level map to a sentence-level one by merging each word into the corresponding sentence node. Design a redundancy layer in the sentence-level diagram which first pre-labels each sentence and renews the label dependencies by spreading the reuse information. The redundancy layer controls the size of the reception field for redundancy information, and the transmission of information is guided by the ground-true labels of the sentences. After receiving a threshold, the whole structure extracts simultaneous abstract sentences instead of the automated progress model, which takes over the top-k strategy. Unlike the previous work, the directed out edge is used to find the similarity and coherent of the sentences that are connected with the previous sentences. The directed in edge, finds the coherent from sentence_1 to sentence_2 whereas the directed out edge finds the coherent from sentence_2 to sentence_1. This ensures the relationship among the sentences more accurately.

3. RESULTS AND DISCUSSIONS

3.1 Datasets

As it is appeared in Table 1, this research work utilizes three datasets generally utilized with various sentences summary that are as follows:

- CNN [13]
- DailyMail [13]
- NYT (New York Times) [14]

In the above datasets, newsroom is purely built for the abstractive summarization. The other datasets were utilized for abstractive summarization.

Table 2 Data Description

S. No.	Datasets	Average document length		Average summary length	
		Words	Sentences	Words	Sentences
1	NYT	801	36	46	3
2	Dailymail	654	30	55	4
3	CNN	761	34	46	4

3.2 Evaluation Metrics

For evaluating summarization technique, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is used.

$$ROUGE-N(c,r)=\frac{\sum_{r_i \in r}\sum_{n-gram \in r_i} Count(n-gram,c)}{\sum_{r_i \in r} numNgrams(r_i)} \quad (3)$$

where N denotes the grams, c is the candidate and r is the reference. In this research work, it reports the following metrics for evaluation of proposed CATSum:

- ROUGE-1,
- ROUGE-2, and
- ROUGE-L

3.3 Parameter Settings

For abstract layer, it separates the named elements (for e.g [Person], [Organization], [Date], [Location], and [Country]) utilizing sapcy, and supplant them by anonymized tokens ([Person_1], [Organization_1], [Date_1], [Location_1], and [Country_1]). This work have attempted to add reliance parse edges and as the results the work didn't show huge advantages. But it shows the attributable to the realities that are as follows:

- The dependent tree is generously a stage consecutive structure, with little headways for unique data;
- The presentation is affected by the precision of the inward stream annotators.

3.4 Baseline Methods

- ABS is the normal architecture with RNN-based encoder and decoder [14].
- PGC augments the standard Seq2Seq attentional model with pointer and coverage mechanisms [15].
- TransformerABS employs Transformer in text summarization [16].
- BART is pre-trained on large unlabeled data and perform excellent performance with Transformer architecture [17].

The Seybold REPORT

ISSN 1533-9211

3.5 Evaluation

The investigation results on four benchmark datasets are appeared in Table 3, 4 and 5. Clearly CATSum nearly beats all the baselines across the majority of the assessment measurements. For abstractive techniques, these variations of transformer perform incredibly with profound structures and huge scope unlabeled corpus.

Table 3 Evaluation of CNN dataset

S.No	Techniques	Rouge – 1	Rouge – 2	Rouge – L
1	BART	46.43	21.96	42.35
2	ABS	44.95	21.35	40.90
3	Transformer ABS	43.69	20.70	39.85
4	PGC	40.70	17.29	36.40
5	CATSum	46.68	24.40	41.89

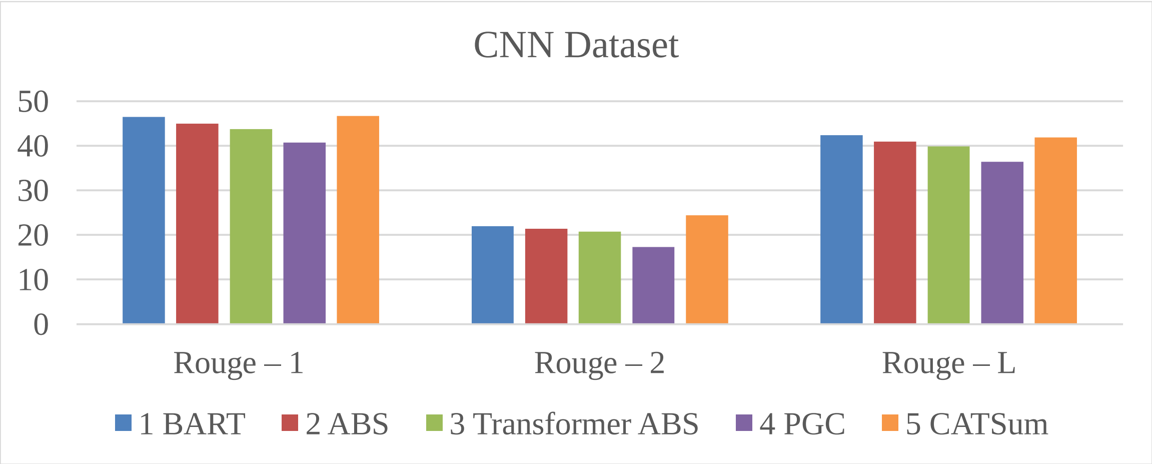

Figure 2 Evaluation of CNN dataset

Table 4 Evaluation of Dailymail dataset

S.No	Techniques	Rouge – 1	Rouge – 2	Rouge – L
1	BART	47.33	22.56	42.35
2	ABS	41.85	22.45	41.70
3	Transformer ABS	41.69	21.60	38.75
4	PGC	40.60	18.39	37.50
5	CATSum	47.68	25.40	42.89

The Seybold Report
ISSN 1533-9211

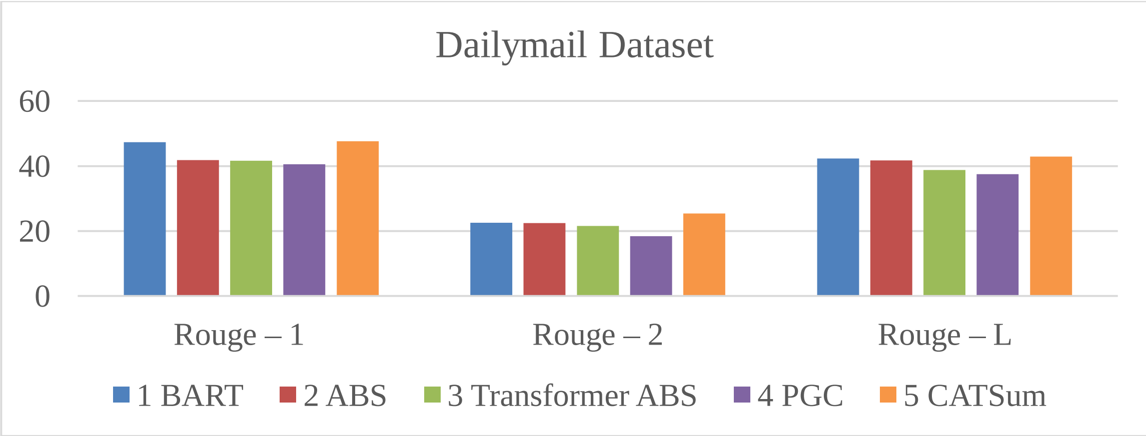

Figure 3 Evaluation of Dailymail dataset

Table 5 Evaluation of NYT dataset

S.No	Techniques	Rouge – 1	Rouge – 2	Rouge – L
1	BART	46.45	23.96	45.45
2	ABS	46.56	24.35	43.20
3	Transformer ABS	47.34	23.70	44.75
4	PGC	42.65	20.29	40.37
5	CATSum	48.78	28.40	46.89

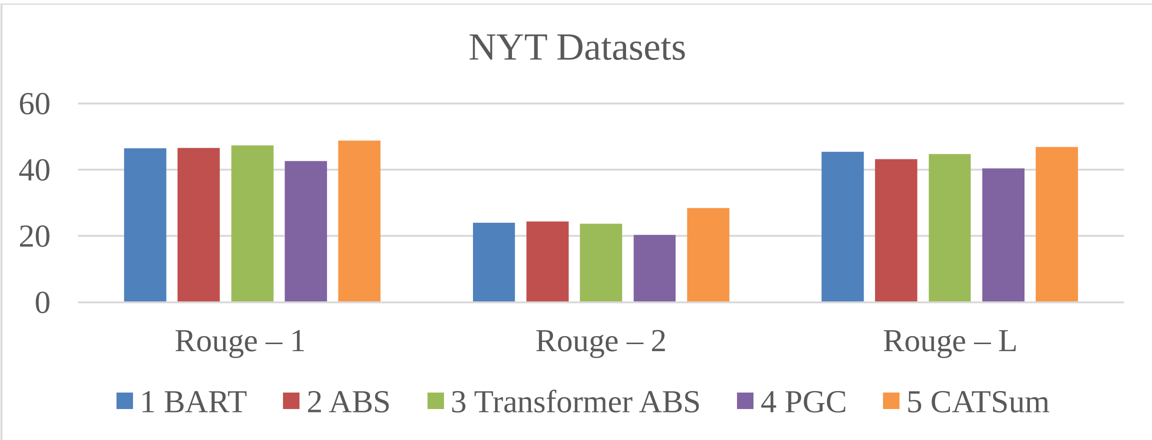

Figure 4 Evaluation of NYT dataset

ISSN 1533-9211

4. CONCLUSION

This research work proposed a categorical attentive different graph. It is targeting salience measure and redundancy concurrently to provide the advance text abstractive summarization. The proposed CATSum approach model did not ignore the redundancy (near-similar) sentences to build the graph. The redundant information provides new scores to evaluate the importance of the sentences. The newly added directed out edge ensure the sentence similarity. The proposed CATSum produces more intensive summaries along with less redundant information.

REFERENCES

1. Akanksha Joshi, E. Fidalgo, E. Alegre, Laura Fernández-Robles, SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders, Expert Systems with Applications, Volume 129, 2019, Pages 200-215,
2. Ge Yao, J., Wan, X., & Xiao, J. Recent advances in document summarization. Knowledge and Information Systems, 53(2), 2017, 297–336.
3. Mihalcea, R., & Tarau, P. A language independent algorithm for single and multiple document summarization. Proceedings of IJCNLP 2005, 2nd International Join Conference on Natural Language Processing (pp. 19–24).
4. Sarkar, K., Saraf, K., & Ghosh, A. Improving graph based multidocument text summarization using an enhanced sentence similarity measure. 2015 IEEE 2nd International Conference on Recent Trends in Information Systems, ReTIS 2015 - Proceedings (pp. 359–365).
5. Tan, J., Wan, X., & Xiao, J. Abstractive document summarization with a graph-based attentional neural model. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, (pp. 1171–1181).
6. Erkan, G., & Radev, D. R. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, volume 22, 2004, 457–479.
7. Kaynar, O., Görmez, Y., Işık, Y. E., & Demirkoparan, F. Comparison of graph based document summarization method. 2017 International Conference on Computer Science and Engineering (UBMK) (pp. 598–603).
8. Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. Banditsum: Extractive summarization as a contextual bandit. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pages 3739–3748.
9. Nallapati, Ramesh, Feifei Zhai, and Bowen Zhou. "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1. 2017.
10. Zhang, Xingxing, Furu Wei, and Ming Zhou. "HIBERT: Document level pre-training of categorical bidirectional transformers for document summarization." *arXiv preprint arXiv:1905.06566* (2019).

The Seybold REPORT

ISSN 1533-9211

11. Narayan, Shashi, Shay B. Cohen, and Mirella Lapata. "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization." arXiv preprint arXiv:1808.08745 (2018).
12. Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pages 3730-3740.
13. *Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2019, arXiv:1909.11942 .*
14. Ming Zhong, Danqing Wang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. A closer look at data bias in neural extractive summarization models, 2019, arXiv:1909.13705.
15. *Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks, 2017, arXiv:1704.04368.*
16. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017, Advances in Neural Information Processing Systems 30 (NIPS 2017).
17. *Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019, arXiv:1910.13461.*