Graph-based Unsupervised Single Document Summarization

Shivali Dubey
Institute for Computational Linguistics
University of Heidelberg
ri197@stud.uni-heidelberg.de

ABSTRACT

Neural network models and availability of large-scale corpus have motivated researchers to achieve new heights in Single Document Summarization. In Supervised Learning, one has to rely on large datasets and at the same time, restricts applications in real-world scenario. Hence, efforts are being made to achieve State-of-the-art results in this field using unsupervised approach. This work is a small attempt at Unsupervised Single Document Summarization using graph-based ranking approach and finding new methods for node centrality computation. In this project, the node centrality of each sentence is computed from its similarity with respect to its neighboring sentences (nodes), to determine how relative position of nodes influence the sentence centrality, hence generating directed graphs. Sentences are represented by word embeddings in the form of tf-idf and BERT embeddings.

1 INTRODUCTION

Till date, a lot of research work has been invested in Supervised Single Document Summarization due to the availability of state-of-the-art neural network architectures and large-scale dataset ([1], [2], [3]). However, in the long run we might not always have abundant data and at the same time we might deal with a totally unrelated dataset. Hence, it is worth experimenting with and studying Unnsupervised Single Document Summarization ([4], [5]). When it comes unsupervised summarization, graph-based approaches have been decently successful. One of the most popular such algorithm is the TextRank ([6])in which sentences (nodes) are connected to each other by undirected graphs, weighted based on sentence similarity. This is where the concept of sentence centrality introduces in PageRank algorithm ([7]) comes into play. The sentences are included in a summary based on their sentence (node) centrality, which is simply a measure of the importance or influence of a given sentence with respect to its neighboring sentences (other sentences within a document). At the same time, sentence centrality is also impacted by the sentimental meaning within a document. Hence, in order to improve the centrality measure BERT ([8]) can be employed.

In this work, (a) the sentence centrality has been estimated using word embeddings in the form of tf-idf and BERT representations. (b) The graphs are constructed such that the edges between sentences (nodes) are directional. It has been observed that most of the times a random ([9]) single sentence in a document would make no sense unless supported by one of the highly weighted central sentence (node). Henceforth, directed graphs take this aspect of graph-based extractive text summarization into consideration.

In this approach, the assumption that the contribution of any two nodes' connection to their respective centrality is influenced by their relative position, is an incentive to measure directed centrality for single-document summarization. The edges are made unidirectional by differentially weighing them according to their orientation. The weights are such that given a pair of sentences in a document, one weighs the edge connecting the following sentence to the preceding sentence and vice versa.

The proposed approach has been evaluated on CNN-Dailymail([10]) news summarization dataset. It has been already shown that ([9]) position-augmented sentence centrality outperforms strong baselines (TextRank, [6]) and this work supports the same. Nevertheless, the achieved results are also comparable to supervised systems trained on thousands of examples. In the past, gensim ([11]), a widely used open-source implementation of TextRank only supported undirected graphs. However, the work ([6]) was taken forward to experiment with position-based directed graphs. It is also worth mentioning that some effort has gone into developing unsupervised models for multi-document summarization ([12]).

2 SENTENCE CENTRALITY MEASUREMENT

Centrality is the measure of the salience of a sentence to be included in a summary, which is prominently used for ranking sentences in graph-based algorithms. As a result, the document is represented as a graph in which nodes are sentences and edges between them are weighted by similarity.The node's centrality can be computed using a ranking algorithm such as PageRank ([7]). This unsupervised approach can be applied by either using directed or undirected graphs. The following sub-sections explain these approaches in detail.

2.1 Undirected Text Graph

For single-document summarization, let D denote a document consisting of a sequence of sentences (s_1, s_2, ...s_n), and $e_{i,j}$, the similarity score for each pair (s_i, s_j). The degree centrality for the sentence s_i can be defined as:

$$Centrality = \sum_{j \in 1,...i-1,i+1,...n} e_{i,j} \quad (1)$$

After obtaining the centrality score for each sentence, sentences are sorted in reverse order and the top ranked ones are included in the summary.Degree centrality only takes local connectivity into account, PageRank ([7]) assigns relative scores to all nodes in the graph based on the recursive principle that connections to nodes having a high score contribute more to the score of the node in question.

2.2 Directed Text Graph

Theories of discourse structure such as Rhetorical Structure Theory (RST, [13]) support the idea that textual units vary in terms of their importance and salience. In RST, elementary discourse units are combined into progressively larger discourse unit, ultimately coverint the entire document. Rhetorical relations link discourse units

which are further characterized in terms of their text importance: *nuclei* denote central segments, whereas *satellites* denote peripheral ones. In this work, the nuclearity is approximated by relative position with an assumption that sentences occurring earlier in a document should be more central. Given any two sentences s_i, s_j ($i < j$) taken from the same document D, this simple intuition is formalized by transforming the undirected edge weighted by the similarity score $e_{i,j}$ between s_i and s_j into two directed ones differentially weighted by $\lambda_1 e_{i,j}$ and $\lambda_2 e_{i,j}$. The centrality score of s_i can be refined based on directed graph as follows:

$$Centrality(s_i) = \lambda_1 \sum_{j<i} e_{i,j} + \lambda_2 \sum_{j>i} e_{j,i} \tag{2}$$

where λ_1, λ_2 are different weights for forward and backward-looking directed edges. When λ_1, λ_2 are equal to 1, Equation (2) becomes degree centrality. The weights can be tuned experimentally on a validation set consisting of a small number of documents and corresponding summaries, or set manually to reflect prior knowledge about how information flows in a document. During tuning experiments, $\lambda_1 + \lambda_2 = 1$ to control the number of free hyper-parameters.

3 SENTENCE CENTRALITY COMPUTATION

The similarity function of TextRank between two sentences can be computed in terms of symbolic sentence representations tf-ids and encoded sentences from BERT ([8]). In this work, unfinetuned BERT and tf-idf representations are subsequently used to compute the similarity between sentences in a document.

3.1 TF-IDF representations

TF-IDF representations are traditional methods in the field of NLP which have been widely used in word embeddings. It is always a good idea to start a project with these representations. In this work too such embeddings have been used to compute sentence similarity.

3.2 BERT as sentence encoder

BERT (Bidirectional Encoder Representations from Transformers; [8]) has been used to map sentences into deep continuous representations. BERT adopts a multi-layer bidirectional Transformer encoder and uses two unsupervised prediction tasks, i.e., masked language modeling and next sentence prediction, to pre-train the encoder.It is a binary classification task, essentially predicting whether the second sentence in a sentence pair is indeed the next sentence. Pre-trained BERT representations can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks. In this work, BERT has been used to encode sentences for unsupervised summarization.

3.3 Similarity Matrix

Once the representations in term of tf-idf and BERT encoded representations ($v_1, v_2, ...v_n$) for sentences ($s_1, s_2, ...s_n$) in document D have been obtained, pair-wise dot product and cosine similarity are employed to compute an unnormalized similarity matrix $\overline{E}$ in terms of dot product:

$$\overline{E} = v_i^T v_j \tag{3}$$

and Cosine Similatity:

$$\overline{E} = \frac{v_i^T v_j}{\|v_i\| \left\|v_j\right\|} \tag{4}$$

Equation(5) aims to remove the effect of absolute values by emphasizing the relative contribution of different similarity scores. This is particularly important for the adopted sentence representations which in some cases might assign very high values to all possible sentence pairs. Hyper-parameter $\beta \in [0, 1]$ controls the threshold below which the similarity score is set to 0.

4 EXPERIMENTAL SETUP

4.1 Dataset

The experiments have been performed on CNN-Dailymail dataset whose statistics are as given below:
Number of documents: 11,490
Average doc. words: 641.9
Average doc. sentences: 28.0
Average summary words: 54.6
Average summary sentences: 3.9
The CNN/DailyMail dataset ([10]) contains news articles and associated highlights, i.e., a few bullet points giving a brief overview of the article. We followed the standard splits for training, validation, and testing used by supervised systems (90,266/1,220/1,093 CNN documents and 196,961/12,148/10,397 DailyMail documents). We did not anonymize entities.

4.2 Implementation details

The items in the validation set were used to tune the hyperparameters ($\lambda_1, \lambda_2, \beta$) on a validation set for both BERT and TF-IDF models. The validation set consisted of 1000 examples with gold summaries. The model performances were further implemented on the test set.

Both fine-tuned and untuned BERT models are available ([8]). I mistakenly have tuned hyperparameters using unfinetuned model, hence the final outputs might vary.

5 RESULTS

5.1 Quantitative Evaluation

Methodically, unigram and bigram overlap (ROUGE-1 and ROUGE-2) are reported as a means of assessing informativeness and the longest common subsequence (ROUGE-L) as a means of assessing fluency. However, in this report I have only reported Rouge-L because the tuning of hyperparameters was performed on the basis of this score. As can be seen from Table 1 that the results are unbelievably good. For BERT model I could say that I used an untuned

Table 1: Quantitative Evaluation.

	Rouge-L
TF-IDF (Dot product)	51.83
TF-IDF (Cosine Similarity)	53.1
BERT	40.96
TF-IDF (PacSum)	35.3
BERT (PacSum)	35.3

model hence the results are not very close to the original work. At the same time I have implemented my own Rouge function, wile in the original work ([9])they used the pyrouge library. Apart from that there were no differences in the implementation. I have also reported the cosine similarity for TF-IDF approach. It can be seen that the Rouge improves upon using Cosine similarity, implying it captures sentence similarity better. In Section(2.2), it was stated that

Table 2: Optimal values of hyperparameters.

	λ_1	λ_2	β
TF-IDF (Dot product)	0.1	0.0	1.0
TF-IDF (Cosine Similarity)	1.0	0.8	0.2
BERT	0.9	0.6	0.4

the optimal λ_1 tends to be lower than λ_2, implying that similarity with previous content actually hurts centrality. From table(2), it can be clearly seen that ony TF-IDF(Dot product) agrees with it. From the tuned hyper-parameters obtained from TF-IDF(Cosine similarity) and BERT, it is more evident the similarity with the following content affects centrality more.

5.2 Qualitative Results

TF-IDF(dot product)
hyp: tennis star caroline wozniacki was forced to defend herself after she congratulated american golf sensation jordan spieth on twitter following his masters success.
ref: caroline wozniacki took to twitter to congratulate golfer jordan spieth.
TF-IDF(Cosine Similarity)
hyp: while the famous kardashian used to pick holidays close to home , in the sun , on a beach and with her family - if her recent vacations are anything to go by - times have changed.
ref: kim and kanye west recently visited armenia to retrace her family 's roots"] hyp.
BERT
hyp: americans of color are disproportionately burdened by the failures of our justice system.
ref: cory booker : the unfortunate reality is that the united states leads the world in incarceration , not education.

6 CONCLUSION

unsupervised summarization system has very modest data requirements and is portable across different types of summaries, domains, or languages. I came across popular graph-based ranking algorithm and understood how node (sentence) centrality is computed. BERT and TF-IDF approaches were applied to capture sentence similarity and build graphs with directed edges arguing that the contribution of any two nodes to their respective centrality is influenced by their relative position in a document. Experimental results do not look very impressive or rather realistic, the reason could be Rouge computation, incorrect hyper-parameter initializations, use of incorrect model(unfinetuned model in case of BERT). In the future this approach could be applied to the task of multi-document-summarization and improve supervised learning approaches. Also, concatenating sentence representations from both BERT and TF-IDF models, also more such embeddings could also give interesting results.

REFERENCES

[1] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. *Summarunner: A recurrent neural network based sequence model for extractive summarization of documents.* In Proceedings of the 31st AAAI Conference on Artificial Intelligence, pages 3075–3081, San Francisco, California. 2017.

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

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

[4] Daraksha Parveen, Hans-Martin Ramsl, and Michael Strube. *Topical coherence for graph-based extractive summarization.* In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1949–1954, Lisbon, Portugal. 2015

[5] Wenpeng Yin and Yulong Pei. *Optimizing sentence modeling and selection for document summarization.* In Proceedings of the 24th International Joint Conference on Artificial Intelligence, pages 1383–1389, Buenos Aires, Argentina. 2015

[6] Rada Mihalcea and Paul Tarau. *Textrank: Bringing order into texts.* In Proceedings of EMNLP 2004, pages 404–411, Barcelona, Spain. 2014

[7] Sergey Brin and Michael Page. *Anatomy of a large-scale hypertextual Web search engine.* In Proceedings of the 7th Conference on World Wide Web, pages 107–117, Brisbane, Australia. 1998

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805. 2018

[9] Hao Zheng and Mirella Lapata. *Sentence Centrality Revisited for Unsupervised Summarization.* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6236–6247 Florence, Italy. 2019.

[10] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt,Will Kay, Mustafa Suleyman, and Phil Blunsom.*Teaching machines to read and comprehend.* In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 1693–1701. Curran Associates, Inc. 2015

[11] Federico Barrios, Federico Lopez, Luis Argerich, and Rosa Wachenchauzer. *Variations of the similarity function of TextRank for automated summarization.* arXiv preprint arXiv:1602.03606. 2016.

[12] Piji Li, Zihao Wang, Wai Lam, Zhaochun Ren, and Lidong Bing. *Salience estimation via variational auto-encoders for multi-document summarization.* In Proceedings of the 31st AAAI Conference on Artificial Intelligence, pages 3497–3503, San Francisco, California. 2017

[13] William C Mann and Sandra A Thompson. *Rhetorical structure theory: Toward a functional theory of text organization.* Text-Interdisciplinary Journal for the Study of Discourse, 8(3):243–281. 1988
https://github.com/mswellhao/PacSum