Computers, Materials & Continua
DOI: 10.32604/cmc.2022.027301
Article

Tech Science Press

Unsupervised Graph-Based Tibetan Multi-Document Summarization

Xiaodong Yan[1,2], Yiqin Wang[1,2], Wei Song[1,2,*], Xiaobing Zhao[1,2], A. Run[3] and Yang Yanxing[4]

[1]School of Information and Engineering, Minzu University of China, Beijing, 100081, China
[2]National Language Resource Monitoring & Research Center, Minority Languages Branch, Beijing, 100081, China
[3]University of California, Irvine, California, 92617, USA
[4]Department of Physics, New Jersey Institute of Technology, Newark, New Jersey, 07102-1982, USA
*Corresponding Author: Wei Song. Email: sw@muc.edu.cn
Received: 14 January 2022; Accepted: 02 April 2022

Abstract: Text summarization creates subset that represents the most important or relevant information in the original content, which effectively reduce information redundancy. Recently neural network method has achieved good results in the task of text summarization both in Chinese and English, but the research of text summarization in low-resource languages is still in the exploratory stage, especially in Tibetan. What's more, there is no large-scale annotated corpus for text summarization. The lack of dataset severely limits the development of low-resource text summarization. In this case, unsupervised learning approaches are more appealing in low-resource languages as they do not require labeled data. In this paper, we propose an unsupervised graph-based Tibetan multi-document summarization method, which divides a large number of Tibetan news documents into topics and extracts the summarization of each topic. Summarization obtained by using traditional graph-based methods have high redundancy and the division of documents topics are not detailed enough. In terms of topic division, we adopt two level clustering methods converting original document into document-level and sentence-level graph, next we take both linguistic and deep representation into account and integrate external corpus into graph to obtain the sentence semantic clustering. Improve the shortcomings of the traditional K-Means clustering method and perform more detailed clustering of documents. Then model sentence clusters into graphs, finally remeasure sentence nodes based on the topic semantic information and the impact of topic features on sentences, higher topic relevance summary is extracted. In order to promote the development of Tibetan text summarization, and to meet the needs of relevant researchers for high-quality Tibetan text summarization datasets, this paper manually constructs a Tibetan summarization dataset and carries out relevant experiments. The experiment results show that our method can effectively improve the quality of summarization and our method is competitive to previous unsupervised methods.

Keywords: Multi-document summarization; text clustering; topic feature fusion; graphic model

1 Introduction

With the development of the mobile Internet media platform, the information on the Internet has exploded. The massive amount of information has brought abundant information to users, but also caused huge reading barriers. In order to meet the needs of users to quickly obtain effective information, text summarization technology emerges as the times require. Automatic text summarization technology uses computer technology to extract content from texts to generate summaries, which can help people quickly obtain information.

Despite recent years have witnessed an increasing number of summarization systems [1,2], but most of those systems are aim to high-resource languages. Low-resource languages summarization systems are still in its infancy. Tibetan is an official language of the Tibet Autonomous Region in China. In addition to China, Tibetan is also spoken in Nepal, Bhutan, and India. The lack of data severely limits the development of Tibetan text summarization. This paper concentrates on survey and realization Tibetan text summarization.

The structure of this research is as follows: Section 2 presents the related work. Section 3 presents our model architecture. Section 4 detailed description about dataset constructions. Section 5 step by step briefing of the proposed technique with tables and graphs. Section 6 implement and evaluation the proposed method. Section 7 conclusion of the entire work.

2 Related Work

Multi-document summarization (MDS) is an effective tool for information aggregation that generates an informative and concise summary from a cluster of topic-related documents [3]. Most of the existing clustering methods directly use k-means to cluster documents. In general, there are two approaches to MDS, extractive approach: words, phrases or sentences are identified as salient pieces of text and reassembled as the summary, does not generate new text; abstractive approach: abstractive summarization approach does not simply copy important phrases from source text but also potentially come up with new phrases, which can be seen as paraphrasing. Recent years variants of neural sequence-to-sequence models have been particularly successful in the summarization tasks [4]. Despite the huge efforts of using deep neural models in summarization, they often require large-scale parallel corpora of input texts paired with their corresponding output summaries for direct supervision [5]. Obtaining training data for MDS is time consuming and resource-intensive, therefore, low-resource languages mostly use unsupervised methods to generate text summaries. According to the selection of features, unsupervised document summarization methods can be divided into the following three approaches: statistics-based approach, topic-based approach and graph-based approach.

Statistics-based approach was first applied to document summarization, which calculates the importance of sentences based on the statistical characteristics of the text to extract summaries. Loret et al. measure the weight of a sentence based on the frequency of the word and the length of the gerund phrase. Statistics-based approach lacks the understanding of the deep semantic relations between sentences, which leads to problems such as difficulty in expressing the topic of the document and logical order missing in summaries [6]. The topic-based approach selects sentences that represent the subject of the document by mining the underlying semantic information of the text. Chang et al. considered the relationship between words, sentences, topics, and documents, and proposed a method to measure weight through the KL divergence between the sentence distribution model and document distribution model [7]. Balaji et al. proposed a method to identify key topics and extract summary from multiple documents [8]. Alrumiah et al. proposed a summarization method by using Latent Dirichlet Allocation (LDA) and length enhancement [9]. The topic-based approach solves the problem of lack of

summarization semantics for a certain extent, but it still lacks information of document structure. The graph-based approach transforms the traditional extraction step into a graph construction, calculation and sorting nodes [10]. Graph-based approaches are widely used in the field of text summarization. The earliest work can be found in [11]. In terms of graph sorting, the classic sorting algorithms include TextRank [12], HITS [13] etc. Most of the existing systems make corresponding improvements based on the text graph constructed by TextRank or HITS algorithms, Li Wei et al. used external corpus information into TextRank in the form of word vectors and use k-means at sentences level to cluster documents [14]. Saeed et al. proposed an abstractive summarization technique which generates variable-length keywords as per document diversity instead of selecting fixed-length keywords for each document, improves the metadata similarity to the original text. [15]. Hu et al. proposed an automatic text summarization technology based on affinity graphs combining topic information to extract highly informative and highly unique sentences [16].

3 Model Architecture

Since the documents come from different sources, the opinions expressed are usually redundant and repetitive. Therefore, the two-level clustering method is adopted. When construct sentence graph considers the deep representation of language together with words embedding. After that, apply spectral clustering to get sentence-clusters, then perform topic feature fusion on each cluster to generate Tibetan summary. In this paper, we propose an unsupervised graph-based Tibetan multi-document summarization method, as shown in Fig. 1. In summary, the contributions of this paper are threefold, as described below:

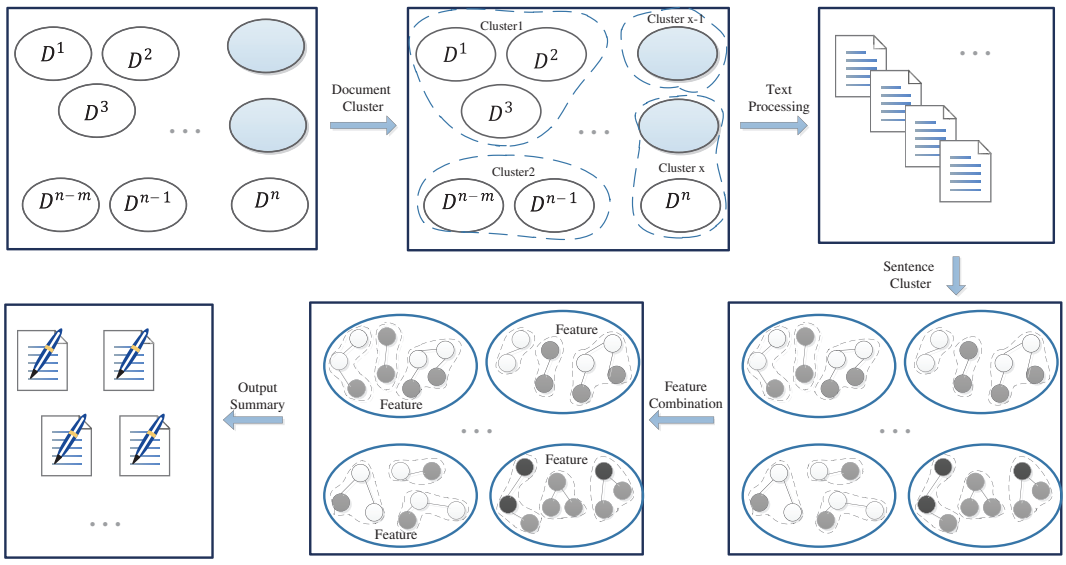


Figure 1: Tibetan multi-document summarization model architecture

1. We introduce an Tibetan multi-document clustering algorithm based on graph model,which two-level clustering methods are performed at the document-level and sentence-level, use two-level clustering can effectively reduce the operating efficiency drop caused by directly constructing sentence graph.

2. We adopt spectral clustering method at sentence level. Define a feature vector for each sentence, then use these sentences features to clustering sentences. Which improve the shortcomings of the traditional K-Means clustering method, multiple documents are divide into finer divisions.
3. We adopt a topic feature fusion method to generate Tibetan text summary, aiming at the traditional graph models for text summarization lack mining and utilization of deep topic semantic features. According to text span and their relevance to input "manual features", reset the weight of the nodes in the graph model. Select the top K nodes with the highest weight in the graph as summary.

4 Dataset Constructions

The construction of datasets is an important task in text summarization. At present, the deep learning algorithm have achieved impressive performance on High-Resource Languages (HRL) datasets, which even surpassing human performance. Such as: DUC dataset, Gigaword dataset and CNN/Dailymail dataset [17,18]. But for low-resource languages, the task of text summarization is still in its infancy due to the lack of corresponding datasets. In order to promote the development of Tibetan text summarization, and to meet the needs of relevant researchers for high-quality Tibetan text summarization datasets. We manual construct a Tibetan text summarization dataset, which contains 1000 parallel of news content paired with their corresponding summaries and more than 3,500 keywords.

4.1 Construction Process

All news in this dataset comes from the "Public Opinion Convergence and Analysis" project of the Natural Language Processing Laboratory of Minzu University of China. First we select the original news, delete those news that are too long or too short. Then clean those texts. Participants in the construction are divided into two groups. One group is responsible for the manual construction of summaries on the cleaned dataset, and the other group is responsible for verifying the quality of the summaries generated above, reviewing the initial summaries, and deleting or manually rebuilding the summaries which below standard.

4.2 News Selection

We adopt 5000 news as initial dataset,those news are crawled from websites such as People's Daily Online, Yunzang Net, Xinhua Net and other websites in the "Public Opinion Convergence and Analysis" project of the Natural Language Processing Laboratory of Minzu University of China. Involving categories such as politics, science and technology, society, economy, art, sports, etc., regular expressions are used to clean text and non-text data such as images, tables, website links, and article sources. In order to improve the quality of the summarization dataset, we discard news texts with less than 1000 words or more than 400 sentences, and finally selected 1,000 news contents for Tibetan summarization dataset construction.

4.3 Summary Construction

The work of summaries constructing are in charge of Tibetan language and literature students from Minzu University of China. Tibetan as their native language, and they also have the basic literacy skills of their major, so they are fully competent in Tibetan summarization writing. The summaries are constructed based on the following requirements: briefly explanation of the materials, highlighting the key points of the news, abandons the content that has no related with the topic. Rigorous sequence

structure and clear hierarchy are necessary. In addition, in order to further improve the quality of the dataset, cross-validation is used to select the constructed summaries.

4.4 Cross-validation

After obtain the initial summarization, the quality of summarizes needs to be verified. The verification group scores the initial summarization from the fluency of sentences, completeness of semantics and coverage of news, and then eliminated low-quality abstracts. The scoring rules are shown in Tab. 1. Remove or rewrite summarization with an average score of less than 3.5. Eventually, 1,000 news and news summarization pairs were manually proofread. Examples of manual construction of summarization are shown in Appendix A.

Table 1: Manual summarization scoring rules

Sentence fluency	influent 1–2	fluent 2–4	clear logical relationships 4–6
Semantic completeness	incomplete 1–2	complete 2–4	complete and easy to understand 4–6
News content coverage	partial 1–2	full coverage 2–4	description based on time and space logic 4–6

5 Model Details

5.1 Graph-based Clustering Algorithm

Text clustering is the application of cluster analysis to text documents. It uses machine learning and natural language processing (NLP) to understand and categorize unstructured, textual data. Through text clustering, texts in the same cluster are more similar to each other than to those in other clusters, so that a set with higher similarity can be founded, and reduce redundancy of text summaries. Text clustering should fully reflect the characteristics of high cohesion and low coupling. Through text clustering algorithm sentences on the same topic can be grouped into same cluster.

Most of the existing clustering methods directly use k-means to cluster document at sentences level. However, due to the large number of sentences in multi-documents, building a sentence-level graph model directly will lead to a decrease in operating efficiency and the topics obtained by directly using K-means to cluster are not detailed enough. Therefore, we adopt two-level of document-level and sentence-level text clustering algorithms to reduce the operating efficiency drop caused by directly constructing sentence graph. First, construct document-level graph model and perform text clustering. Secondly, construct a sentence graph for the sentences of the documents in the obtained clusters, and assign a feature vector to each sentence, then perform clustering on these feature vectors.

5.1.1 Document-level Clustering

The news corpus needs to be clustered by topic firs in order to generate multi-document summaries for the documents under each topic. The typical clustering algorithm is the K-Means clustering algorithm [19]. Since K-Means is an unsupervised algorithm and does not require a training set, it can effectively save clustering costs, so K-Means has become one of the most widely used clustering

algorithms [20]. We use K-Means combined with semantic similarity method for document clustering. Based on the text vector space model, the cosine similarity is used to calculate the similarity between two documents, and document-level graph model is constructed according to the similarity threshold of each document.

5.1.2 Sentence -level Clustering

After the document-level graph model is constructed, document-level text clustering algorithms will be used to obtain document clusters with high similarity, that is, the discovery process of subtopics. In order to divide a topic in more details, the sentences under the subtopics are clustered.

In terms of sentence graph construction, which is different from document graph construction, we construct a sentence graph based on Approximate Discourse Graph (ADG) [21]. Specifically, we build a graph (V, E), where each node $v_i \in V$ represents a sentence, and nodes v_i and v_j $(i \neq j)$ are connected, i.e., their edge $e_{i,j} = 1$ if the similarity between sentences is greater than the threshold. Sentence level clustering schematic diagram as shown in Fig. 2.

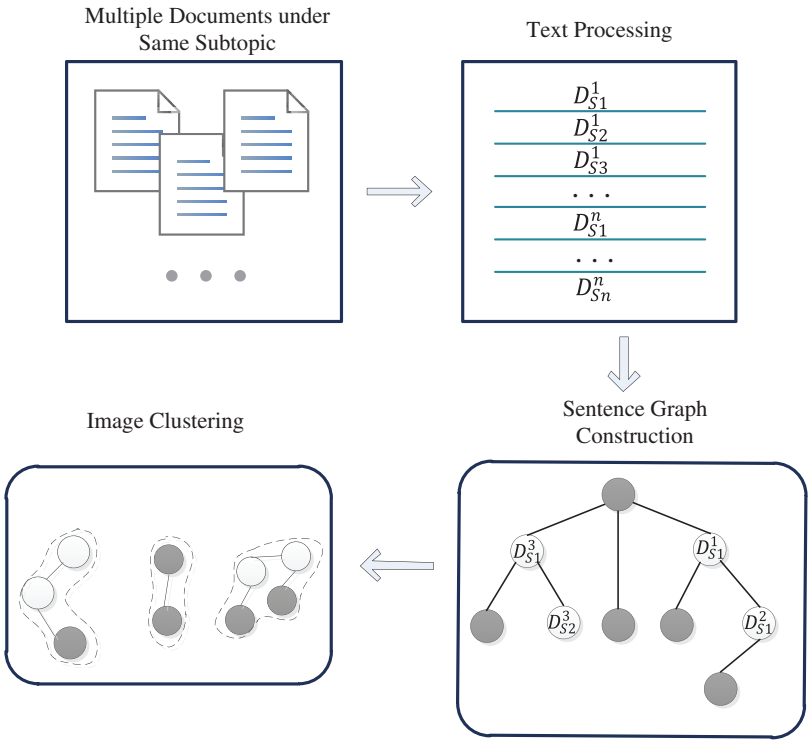


Figure 2: Sentence clustering diagram

K-means is sensitive to the initial clustering center, and the division of dense dataset such as text is not detailed enough [22]. For sentence level cluster, we use spectral clustering method. The clusters obtained by the spectral clustering method have the characteristics of small intra-cluster distances and large inter-cluster distances, enable more detailed topic division of documents.

5.2 Tibetan Text Summarization Combining Topic Feature

Our Tibetan text summarization combine topic feature method is a graph-based content extraction method inspired by the TextRank algorithm. We use keywords which were manually generating a general description of the document as the topic of news. Reassign the random restart probabilities of graph nodes based on the relevance of the graph nodes to the topic of the news, enable sentences that related to the topic have higher score.

5.2.1 TextRank Node Scoring

In the original TextRank algorithm, the strengths and weaknesses of text spans reflect through links extracted directly from the original text. TextRank treat each sentence in the text as a graph node v_i and the edges between nodes having a weight $w_{i,j}$, $w_{i,j}$ is calculated by the similarity between sentence nodes, sentence similarity score is calculated by taking cosine similarity of two sentence vectors, sentence vectors are obtained by averaging all of the word vectors of a sentence. TextRank node score as shown in Eq. (1):

$$TextRank\,(V_i) = (1-d) + d* \sum_{v_j \in (v_i)} \frac{w_{i,j}}{\sum_{v_k \in Out(v_j)}} TextRank(V_j) \tag{1}$$

5.2.2 Feature Combining Node Scoring

In the traditional TextRank algorithm, each node has an equal random restart probability, so all nodes are treated equally during the application of the algorithm. However, we hope that the higher the relevance of the sentence to the topic of document, the higher the probability that the sentence will be selected. We reset the node score based on Biased-TextRank [23] combining topic feature. The TextRank node score combine with the topic feature is shown in Eq. (2):

$$TextRank\,(V_i) = \text{feature}^*\,(1-d) + d* \sum_{v_j \in (v_i)} \frac{w_{i,j}}{\sum_{v_k \in Out(v_j)}} TextRank(V_j) \tag{2}$$

Among them, the feature value is set to reflect the relevance of the current sentence node and the topic keyword, and the damping factor d is set to 0.85 as described above. We use multiple keywords extracted from the description content to determine the similarity between nodes and topics. Convert multiple keyword information into fixed-length embedding vectors, and calculate the similarity with nodes. The higher the similarity between the node and the embedding vector, the higher the restart probability assigned to the node. Finally, the first K nodes with high weight are selected as the summary sentence, and K is selected as 20% of the length of the article.

6 Experiments

6.1 Data Preprocessing

We use Tibetan border characters to separate sentences. Then remove the stop words and punctuation through the Tibetan stop words list, and use the TIP-LAS [24] tool to segment words. Consider that sentences that are too long or too short are not suitable as candidate sentences for the abstract, and those that are too long or too short are removed.

6.2 Evaluation Method and Dataset

Evaluation methods are a key part of the task of text summarization. The evaluation methods of text summarization can be roughly divided into two categories: Intrinsic Methods and Extrinsic

Methods. Internal method: Provide reference summary, and evaluate the quality of the system summary based on the reference summary. The quality of the system is evaluated by the degree of agreement between the system summary and the reference summary. External evaluation methods do not provide reference summary and are generally applied to specific tasks. For example: document retrieval, document clustering, document classification, etc., to evaluate the quality of summary based on whether the summary can improve application performance. Among them, the internal method is the most commonly used summary evaluation method in academia. Comparing the system summaries with the expert summaries using a certain method is also one of the most common summary evaluation methods at present. The expert summaries are used as reference summary to evaluate the quality of system summary. Lin et al. [25]. proposed the ROUGE automatic summary evaluation method based on BLUE, an automatic evaluation method for machine translation, which is now widely used in summary evaluation tasks. ROUGE compares expert summary with system summary, counts the overlapping basic units, and evaluates the quality of the system summary. At present, ROUGE has become one of the general standards for summary evaluation. ROUGE is an evaluation method for the recall rate of n-grams,the calculation shown as Eq. (3):

$$ROUGE-N=\sum_{v_j\in(v_i)}\frac{\sum_{S\in Refsummarizes}\sum_{n-grams\in S}Count_{match}(n-gram)}{\sum_{S\in Refsummarizes}\sum_{n-grams\in S}Count(n-gram)} \tag{3}$$

Where RefSummaries represents reference summaries, that is, expert summaries obtained in advance, $Count_{match}(n-gram)$ represents the number of co-occurrences of n-grams of system summary and reference summary, and $Count(n-gram)$ represents the number of n-grams that appear in the reference summary.

Since there is no general dataset in the field of Tibetan text summarization research, we first use news as corpus and title as reference summary for evaluation [26]. In order to verify the performance of the system summary on the expert summary dataset, we use the Tibetan summary dataset, using news as the corpus, and expert summary as the reference summary for evaluation. Then compare the evaluation results of the two methods.

6.3 Experimental Results and Analysis

The ROUGE evaluation results of the news title as reference summary are shown in Tab. 2. The key sentences extracted by our method have the best effect on the ROUGE evaluation index where the title is used as reference summary. However, when the news headline is used as the reference summary, the ROUGE score of each method is relatively low. This is because the headline of the news usually only has one or two sentences, so the headline only summarizes the parts of news content, and lack of a comprehensive description of news events which cannot provide a complete summary of news. Therefore, using the title as a reference summary cannot evaluate the comprehensiveness of the system summary. The Tibetan summarization dataset was constructed, and the effect of system summary was evaluated on the Tibetan summarization dataset.

Table 2: Rouge evaluation results of title reference summary

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead-3+K-Means	10.9	6.5	10.9
TextRank+ K-Means	12.5	5.7	11.1

(Continued)

Table 2: Continued

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead-3+ two-level clustering	14.5	9.6	13.3
TextRank+ two-level clustering	18.1	8.8	16.3
Our method	32.4	18.5	25.9

The Tibetan summarization dataset constructed in this paper can provide a concise and comprehensive summary of news content. We use this dataset as reference summary for evaluation, use ROUGE as an evaluation indicator, and conduct the following experiments.

Lead-3+ K-Means: We use K-Means combine with Lead-3 as the baseline of the experiment.

TextRank+ K-Means: We use K-Means combine with TextRank to extract summaries. Use word frequency co-occurrence matrix to calculate similarity.

Lead-3+ two-level clustering: We use two-level clustering combine with Lead-3 to extract abstracts to verify the effectiveness of two-level clustering.

TextRank two-level clustering: Use two-level clustering combine with TextRank method to extract summary. The word frequency co-occurrence matrix is used to calculate the similarity.

The expert summary used as the reference summary for evaluation. The evaluation results of ROUGE-1, ROUGE-2, and ROUGE-L are shown in Tab. 3.

Table 3: Rouge evaluation results of manual reference summary

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead-3+K-Means	17.1	10.1	15.4
TextRank+ K-Means	19.5	9.2	18.2
Lead-3+ two-level clustering	20.6	11.6	17.7
TextRank+ two-level clustering	26.6	8.4	21.3
Our method	33.9	27.1	32.6

When use lead-3 method as the experimental baseline, our method compared with the baseline, ROUGE-1 increased by 16.8%, ROUGE-2 score increased by 17%, ROUGE-L score increased by 17.2%. Our method compared with the K-Means + TextRank method, ROUGE-1 score increased by 14.4%, ROUGE-2 score increased by 17.9%, and ROUGE-L score increased by 14.4%, which proved the effectiveness of two-level clustering and topic feature fusion. Compared with the two-level clustering + TextRank method, the ROUGE-1 score has increased by 7.3%, ROUGE-2 score has increased by 18.7%, and ROUGE-L score has increased by 11.3%, which verified that the method of topic feature combination can generate a summary more in line with the topic.

As shown in Fig. 3 the score ranking of various methods in the title reference summary is basically the same as the score of the expert reference summary. However, the ROUGE scores have improved on the expert reference summary. This is because Tibetan summarization dataset can achieve a comprehensive and focused evaluation of summary, also comprehensively consider the results of the system summary. Our method is more optimized than traditional algorithms in Tibetan

multi-news summarization. Use two-level graph model for multi-text clustering, when processing high-dimensional data such as vectors, the complexity of clustering is better than traditional clustering algorithms. The text summarization combine topic feature method can select candidate sentences that are more relevant to topic. Through the method of graph model clustering & topic feature fusion the obtained summary can describe the news content comprehensively.

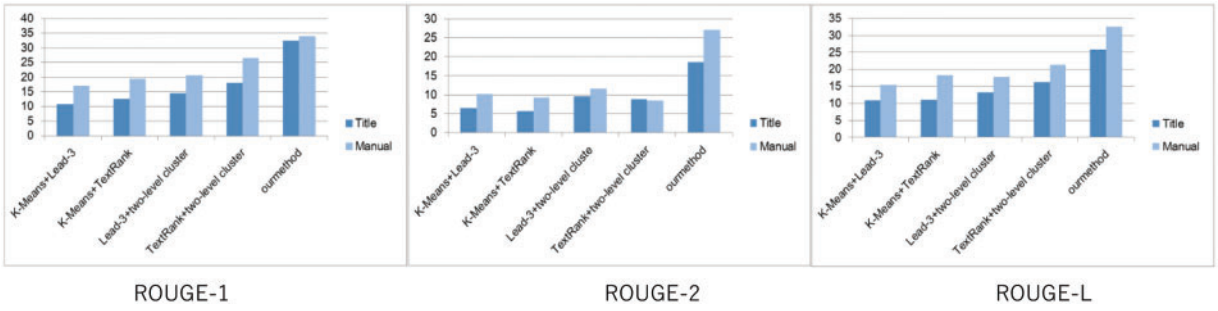


Figure 3: Score comparison

7 Conclusions

We propose an unsupervised Tibetan multi-document summarization method based on graph model. The two-level clustering can effectively improve the efficiency of the algorithm, and the generated summaries are more hierarchical. Based on the topic semantic information reflects the main idea of the news and the impact of topic features on sentences, the value of sentence nodes in the graph is re-measured. The method of topic features combine with summary extraction is proposed. By manually construct a Tibetan summarization dataset, the experiment of extracting Tibetan summarization on the dataset has achieved good results, the effectiveness of the Tibetan summarization method that proposed in this paper has been verified. Since the graph model method used in this paper is an unsupervised algorithm, we did not use a large-scale corpus in the experiment, which has certain limitations. In the next step, we will expand the scale of the Tibetan summarization dataset and try to generate abstractive summary on the large-scale Tibetan summarization dataset.

Funding Statement: This work was supported in part by the National Science Foundation Project of P.R. China 484 under Grant No.52071349, partially supported by Young and Middle-aged Talents Project of the State Ethnic Affairs 487 Commission.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Hu, X. Li, Y. Deng, Y. Peng, B. Lin *et al.,* "A semantic supervision method for abstractive summarization," *Computers Materials & Continua*, vol. 69, no. 1, pp. 145–158, 2021.

[2] E. Heidary, H. Parvïn, S. Nejatian, K. Bagherifard, V. Rezaie *et al.,* "Automatic text summarization using genetic algorithm and repetitive patterns," *Computers Materials & Continua*, vol. 67, no. 1, pp. 1085–1101, 2021.

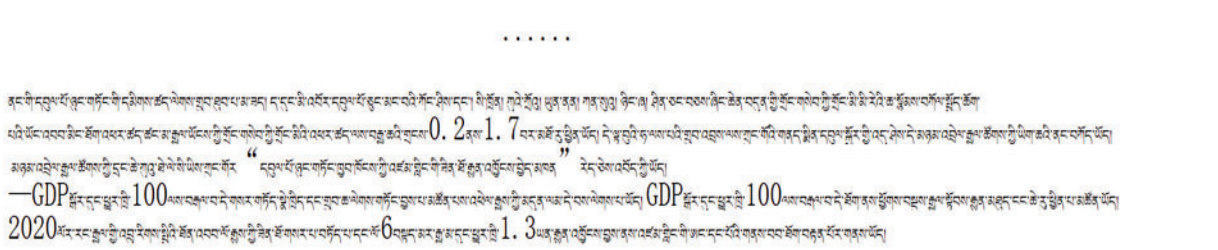
[3] C. Ma, W. E. Zhang, M. Guo, H. Wang and Q. Z. Sheng, "Multi-document summarization via deep learning techniques: A survey," *arXiv Preprint*, arXiv.2011.04843, 2020. [Online]. Available: https://arxiv.org/abs/2011.04843.

[4] R. Nallapati, B. Zhou, C. Gulcehre and B. Xiang, "Abstractive text summarization using sequence-to-sequence rnns and beyond," *arXiv Preprint*, arXiv.1602.06023, 2016. [Online]. Available: https://arxiv.org/abs/1602.06023.
[5] J. Zhao, M. Liu, L. Gao, Y. Jin, L. Du *et al.,* "Summpip: Unsupervised multi-document summarization with sentence graph compression," in *Proc. SIGIR*, Xi'an, China, pp. 1949–1952, 2020.
[6] E. Lloret and M. Palomar, "A gradual combination of features for building automatic summarization systems," in *Proc. TSD*, Pilsen, Czech Republic, pp. 16–23, 2009.
[7] Y. L. Chang and J. T. Chien, "Latent Dirichlet learning for document summarization," in *Proc. IEEE.ICASSP*, Taipei, Taiwan, China, pp. 1689–1692, 2009.
[8] J. Balaji, T. V. Geetha and R. Parthasarathi, "A Graph based query focused multi-document summarization," *International Journal of Intelligent Information Technologies*, vol. 10, no. 1, pp. 16–41, 2014.
[9] S. S. Alrumiah and A. A. Al-Shargabi, "Educational videos subtitles' summarization using latent dirichlet allocation and length enhancement," *Computers, Materials & Continua*, vol. 70, no. 3, pp. 6205–6221, 2022.
[10] K. S. Thakkar, R. V. Dharaskar and M. B. Chandak, "Graph-based algorithms for text summarization," in *Proc. ICETET*, Washington, DC, United States, pp. 516–519, 2010.
[11] I. Mani and E. Bloedorn, "Multi-document summarization by graph search and matching," *arXiv Preprint*, arXiv. cmp-lg/9712004, 1997. [Online]. Available: https://arxiv.org/abs/cmp-lg/9712004.
[12] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in *Proc. EMNLP*, Barcelona, Spain, pp. 404–411, 2004.
[13] Y. Sankarasubramaniam, K. Ramanathan and S. Ghosh, "Text summarization using Wikipedia," *Information Processing & Management*, vol. 50, no. 3, pp. 443–461, 2014.
[14] W. Li, X. D. Yan and Q. Xie, "An improved Textrank for Tibetan summarization," in *Proc. CCL*, Hainan, Haikou, China, pp. 36–43, 2020.
[15] M. Y. Saeed, M. Awais, M. Younas, M. A. Shah, A. Khan *et al.,* "An abstractive summarization technique with variable length keywords as per document diversity," *Computers Materials & Continua*, vol. 66, no. 3, pp. 2409–2423, 2021.
[16] P. Hu, J. He and Y. Zhang, "Graph-based query-focused multi-document summarization using improved affinity graph," in *Proc. KSEM*, Chongqing, China, pp. 336–347, 2015.
[17] K. C. Litkowski, "Summarization experiments in DUC 2004," in *Proc. HLT-NAAC*, Boston, MA, USA, pp. 6–7, 2004.
[18] C. Napoles, M. R. Gormley and B. VanDurme, "Annotated gigaword," in *Proc. AKBC-WEKEX*, Montreal, Canada, pp. 95–100, 2012.
[19] A. Likas, N. Vlassis and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451–461, 2003.
[20] C. C. Aggarwal and C. Zhai, *A survey of text clustering algorithms*. Boston, MA: Mining text data Springer, pp. 77–128, 2012.
[21] J. Christensen, S. Soderland and O. Etzioni, "Towards coherent multi-document summarization," in *Proc. NAACL*, Atlanta, United States, pp. 1163–1173, 2013.
[22] M. Arun and S. Swamynathan, "Hierarchical stream clustering based news summarization system," *Computers Materials & Continua*, vol. 70, no. 1, pp. 1263–1280, 2022.
[23] A. Kazemi, V. Pérez-Rosas and R. Mihalcea, "Biased TextRank: Unsupervised graph-based content extraction," *arXiv Preprint*, arXiv.2011.01026, no. 2, 2020. [Online]. Available: https://arxiv.org/abs/2011.01026.
[24] B. H. Li, H. D. Liu, C. J. Long and J. Wu, "Tibetan word segmentation based on deep learning," *Computer Engineering & Design*, vol. 39, no. 1, pp. 194–198, 2018.
[25] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. ACL-04 Workshop*, Barcelona, Spain, pp. 74–81, 2004.
[26] C. Napoles, M. R. Gormley and B. Van Durme, "Annotated gigaword," in *Proc. AKBC-WEKEX*, Montreal, Canada, pp. 95–100, 2012.

Appendix A.

Algorithm A: Examples of the Tibetan summarization dataset

original Tibetan text



.....

(By 2020, China's gross domestic product (GDP) will exceed 100 trillion yuan for the first time, reaching 101,5986 billion yuan. Calculated at comparable prices, it will increase by 2.3% over the previous year. GDP will exceed 100 trillion yuan, and we will work together and share Development is integrated and everyone has income. Yunnan proposes that all the Lisu, Dulong, Nu and Pumi populations will be lifted out of poverty, and all 88 impoverished counties will be lifted out of poverty. Guangxi requires that 8 impoverished counties including Rongshui Miao Autonomous County have their hats removed, and 54 impoverished counties Decapitated. Guizhou Province announced that nine poverty-stricken counties in the province have exited poverty-stricken counties. China's economy has reached a new level. According to preliminary calculations, China's gross domestic product (GDP) will exceed 100 trillion yuan for the first time in 2020, reaching 1015986 100 million yuan. Calculated at comparable prices, an increase of 2.3% over the previous year. -GDP exceeded 100 trillion yuan, hard-won and very rare. This reflects the central government's ability to judge and make decisions. By 2020, a hundred years the unprecedented SARS epidemic came suddenly, and the world economy was in the most severe ups and downs since the end of World War II. Faced with the sudden shock, China's GDP in the first quarter fell by 6.8% year-on-year, achieving the first negative growth since quarterly statistics. "Only fully demonstrated. Only with the huge potential and powerful driving force of our country's development can we achieve the goals and tasks of this year's economic and social development."

. . . . . .

(Continued)

Summarization

ཀྲུང་གོའི་དཔལ་འབྱོར་སྐས་རིམ་གསར་བ་ཞེག་ཏུ་སྐེབས་པ་སྟེ།

2020ལོར་རང་རྒྱལ་གྱི་རྒྱལ་ནང་ཐོན་སྐྱེད་ཀྱི་རིན་ཐང་བསྡོམས་འབོར(GDP) སྒྲོར་དང་ཕྱར་ཁྲི་100ལས་ཕེབས་དང་དེ་བརྒལ་ནས་སྒྲོར་དང་ཕྱར་1015986ཟེན་ཡོད།

བསྒར་གཞི་ཡོད་པའི་རིན་གོང་ལྟར་རྩིས་ན། ལོ་གོང་མ་ལས་2.3%འཕར་ཡོད།

གོ་ལ་ཉིལ་ཡོའི་ནང་དཔལ་འབྱོར་དུང་འཕར་ཤུབ་པའི་དཔལ་འབྱོར་བྱེད་པོ་གཙོ་བོ་འབའ་ཞིག་དེར་འགྱུར་རྒྱ་དང་འཛོམ་གྲྭའང་གི་དཔལ་འབྱོར་ཁྱོད་ཀྱི་ཐོབ་སྐལ་ཡང་2019ལོའི་16.3%

ནས་17%ཙམ་བར་འཕར་ནས་ལོ་རྒྱས་ཀྱི་ཟེན་ཐོ་གསར་བ་བགོད་ཡོད་པས། ཀྲུང་གོ་གསར་པའི་ལོ་རྒྱས་ཐོག་སྒེར་བཏང་གཏན་ནས་མེན་པའི་ལོ་འདིའི་ནང་མི་དམངས་སྤྲོ་ཡིད་ཚོམ་པ་དང་། འཛམ་གླིང་ས་གོ་ཀན་ཁྱིས་དོ་སྣང་བྱེད་པ། ལོ་རྒྱས་དེའ་ཞེར་ཐོག་འགོད་ཚོག་པའི་རྒྱགས་ལན་ཞིག་ཕྱལ་ཡོད།

The Chinese economy has reached a new level. By 2020, China's gross domestic product (GDP) will exceed 100 trillion yuan for the first time, reaching 101,5986 billion yuan. Calculated at comparable prices, an increase of 2.3% over the previous year. The global economy will become the only economic entity with fair growth, and its share in the world economy will rise from 16.3% in 2019 to around 17%, a record high. An extraordinary year in the history of New China, the people were satisfied. It has attracted worldwide attention. Handed over the answer sheet recorded in the annals of history.