Research on Logistics Text Information Abstract Generation Method

Yao Ma*, Yufeng Zhuangb, Yanzhu Hu*, Wenjia Tianc and Song Wanga

School of Automation, Beijing University of Posts and Telecommunications, Beijing, China

*Corresponding author e-mail: yzhu@263.net, *mayaoheleena@163.com, bzhuangyf@bupt.edu.cn, ctianwenjia@bupt.edu.cn, dwongsangwongsang@163.com

Abstract. The sensational headline news has seriously affected the news browsing experience of netizens, and has caused an increasingly bad impact on the entire network. Due to this problem, the paper proposed a summary generation method based on Graph2Seq model. This method integrates keywords obtained from graph neural network training with contextual information, and integrates them into the Graph2Seq model as an attention mechanism. It plays a central role in guiding the abstract generation process and improves the accuracy of abstract generation.

1. Introduction

With the advent of the big data era, the logistics text information on the Internet has grown exponentially. There are various forms of logistics text information, including messages, information, documents, materials, data, etc. generated and used during the logistics activities. Logistics industry news is also a form of logistics text information, which emphasizes timeliness and authenticity. The headline of logistics news is the highlight of a news report and the basic summary of the true content of the news. Nowadays, the Internet is full of a lot of 'junk' news headlines, and the news we call the sensational headline news are endless. The sensational headline news content generally over-exaggerates or even fabricates the topic in order to attract users' attention. This phenomenon has seriously affected the news browsing experience of netizens, and has caused an increasingly bad impact on the entire network. In order to avoid this situation, it is an urgent need to read the entire news content in advance and provide important information to improve reading efficiency and ensure reading quality.

At present, there are two main methods for automatic generation of text summaries, one is extractive and the other is generative. Generative abstraction is to let the model learn the meaning expressed in the original text, infer the summary content that needs to be generated, and the model automatically generates a summary text. Most of the researches on generative abstracts at present are based on the sequence-to-sequence model (Sutskever et al., 2014) proposed by the GoogleBrain team in 2014. This method is also known as the encoder and decoder architecture. Due to the existence of the "long-distance dependency" problem, the previous sequence-to-sequence method lost a lot of information when encoding the generated speech vector C, resulting in an insufficiently accurate summary. (Bahdanau et al., 2014) firstly applied the attention mechanism to neural networks in NLP and its translation tasks (NMT). The purpose is to make the network always focus on specific content at a certain moment and selectively ignore other content. Facebook AI Research (FAIR) proposed a convolutional network (CNN)
(Gehring et al., 2016) for the Encoder part. The experimental results reached the highest level in the same year in the generation of abstracts. The Google team published a paper on the Attention mechanism (Vaswani et al., 2017), using only Self-Attention and Encoder-Decoder Attention, the end-to-end translation task was fully realized. (Rush et al., 2017) applied the seq2seq model to the summary generation task. Graph neural networks are more and more widely studied and used because they can process and calculate data in non-Euclidean domains and can deal with the complexity of graph data. Many studies have obtained variants of graph neural networks based on different features. (Niepert et al., 2016) proposed a convolutional neural network for graph structures, namely, the graph convolutional network (GCN). (Yao et al., 2019) used graph convolutional network (GCN) to construct and encode graph structure for text, and realized text classification through Text GCN algorithm. (Xu et al., 2018) proposed a new attention-based graph sequence learning neural network model, Graph2Seq, for learning expressive node embeddings and recombining them into corresponding graph embeddings.

In order to solve the problem that the sensational headline news in logistics leads to the reading and obtaining correct logistics news trends, this paper intends to use a graph neural network-based keyword extraction method, attention mechanism generation, and Graph2Seq-based summary generation method to study specific issues. The obtained logistics news text is sent to implement text vectorization to obtain the word vectors. The obtained word vectors and its semantic information are used to generate graph structure data. Then the graph structure data is sent to a graph neural network for training. And the training results are combined with the TextRank algorithm to calculate node weights. The weights are used to extract the keywords. After all of these, the keywords and text graph structure data are integrated into the Graph2Seq model as an attention mechanism. The summary generated by the Graph2Seq model is compared with the news headline to identify the sensational headline news.

2. Summary generation model

The thesis starts from the chaos of logistics sensational headline news and aims to solve the misunderstanding of the sensational headline news and improve the efficiency of news reading. The paper proposed a summary generation method based on Graph2Seq model. This method integrates keywords obtained from graph neural network training with contextual information, and integrates them into the Graph2Seq model as an attention mechanism. It plays a central role in guiding the abstract generation process and improves the accuracy of abstract generation. The framework of the algorithm is shown as follows.

![Figure 1. Framework diagram of abstract generation method in this paper](image-url)
2.1. Keyword Extraction Method Based on Graph Neural Network

The existing TextRank algorithm (Mihalcea et al., 2004) is based on a lexical graph model, because the method is an unsupervised extraction algorithm, the important semantic information of the words and the overall information of the document set are ignored. This paper uses graph neural networks to train logistics news text information to obtain weights, and combines the new weights with the TextRank algorithm to improve the accuracy of logistics news text keyword extraction, so that keywords can be closely related to the logistics theme. TextRank formula is shown below.

\[ S(V_i) = (1-d) + d \sum_{j \in \text{in}(V_i)} \frac{1}{\text{Out}(V_j)} S(V_j) \]  

(1)

2.1.1. Syntactic Analysis and Text Vectorization. The paper firstly needs to perform a syntax analysis on text to obtain the composition and context. The word2vec model represents text as a vector which can express the semantics of the text and ensures the similarity of words as well as in spatial distribution.

2.1.2. Word Vector Generation Graph Structure Data. The word vectors generated by the word2vec model constitute a new corpus, where the dimensionality of the corpus is the dimension set during training, and the number of vectors contained in the corpus is the number of words input during training. Set each word vector in the corpus as a node in the graph data, construct the edges between the nodes according to the semantic dependencies, and finally generate a full-corpus-oriented graph for the first layer of embedding in graph neural network training.

2.1.3. Graph Neural Network Training. The concept of graph neural network (GNN) was first proposed in (F. Scarselli et al., 2009). It extends the existing neural network to process the data represented in the graph domain. The goal of GNN is to learn a state embedding that contains the neighbourhood information of each node, update the state by iteration, and calculate the gradient of weight W from the loss after the state is stable, and then update the weight W according to the gradient. The state iteration formula is as follows:

\[ h_v = f \left( x_v, x_{\text{co}v}, h_{\text{ne}v}, x_{\text{ne}v} \right) \]  

(2)

\[ x_v, x_{\text{co}v}, h_{\text{ne}v}, x_{\text{ne}v} \] are the characteristics of node v, the characteristics of edges, the state of adjacent nodes, and the characteristics of adjacent nodes.

2.1.4. Combination of Keyword Weight and TextRank Algorithm. TextRank algorithm is a mature algorithm that extracts keywords currently. Its principle comes from Google's PageRank algorithm. This paper combines the TextRank algorithm with the training weights of graph neural networks to improve the accuracy of the keyword extraction for logistics news texts and ensure that the extracted keywords are closely related to the logistics theme.

2.2. Attention Mechanism Applied to Abstract Generation

Attention mechanism has been widely used in the field of natural language processing, which is broadly interpreted as a vector of importance weights, that is, in order to predict an element, such as a word in a sentence, the attention vector is used to estimate how strong the relevant it is to other elements. The sum of its values is an approximation of the target. In this paper, the topic keywords obtained through graph neural network training are integrated into the model as an attention mechanism. At the same time, the input and output attention mechanisms are added and jointly embedded in the current output sequence of the decoder. The abstract can also avoid the redundancy and duplication of information by reviewing the output sequence information.
2.3. Graph2Seq-based Summary Generation Method

The sequence-to-sequence (Seq2Seq) model can also be called the Encoder-Decoder framework, which is a general processing model suitable for processing one text to generate another. Although the Seq2Seq model is very flexible and expressive, it can only be applied to problems where the input is a sequence. The Graph2Seq model can be used to process complex structure graph input. It is a new attention-based neural network structure for graph-to-sequence learning and a general end-to-end neural codec structure. Graph2Seq can be regarded as a generalized graph input Seq2Seq model, which uses an encoder-decoder architecture similar to Seq2Seq, including a graph encoder and a sequence decoder. The graph encoder section learns node embedding by aggregating adjacent information in directed and undirected graphs. Then it builds a graph embedding based on the learned node embeddings. The part of the sequence decoder uses an attention-based LSTM network and uses graph embedding as the initial hidden state to output the target prediction. In this paper, the Graph2Seq model is applied to the abstract extraction of logistics news, and the corresponding attention mechanism is added to the graph embedding part to improve the accuracy of generating abstract.

3. Experiment

3.1. Experimental Setup

Keyword Extraction. The data set used in the keyword extraction experiment is a collection of 500 abstracts in the Inspec database, and the corresponding manually assigned keywords. This test data set is the same as the test data set used in the keyword extraction experiment in (Mihalcea et al, 2004). This paper uses a total of a 1000 abstracts for training, of which 700 were used as training sets and 300 were used as test sets.

Abstract Generation. The experimental data used for the abstract generation section included 500 news articles from the New York Times and the Associated Press, each of which was paired with 4 artificially generated reference abstracts (non-headings), with a maximum of 75 words per article Section. The data set of the evaluation part is DUC-2004. ROUGE-1, ROUGE-2, and ROUGE-L were used as evaluation indicators.

3.2. Experimental Results

Keyword Extraction. The experimental results based on graph neural network (GNN) keyword extraction algorithm and TextRank algorithm are shown in Table 1. We use accuracy and recall as the evaluation indicators.

| Number of keywords | index | 3     | 5     | 10    |
|--------------------|-------|-------|-------|-------|
| TextRank Algorithm | Precision | 28.1  | 28.2  | 29.7  |
|                    | Recall  | 37.6  | 37.7  | 39.9  |
| GNN-based extraction algorithm | Precision | 29.1  | 3.0.0 | 31.2  |
|                    | Recall  | 37.8  | 38.1  | 43.1  |

It can be seen from the experimental results that the recall and accuracy of the keyword extraction algorithm based on GNN are improved compared to the TextRank algorithm, which shows that the method of applying GNN to keyword extraction is feasible.

Abstract Generation. The Graph2Seq-based summary generation method uses DUC-2004 as the evaluation data set, and uses ROUGE-1, ROUGE-2, and ROUGE-L as evaluation indicators. The experimental results are compared with some existing models. The experimental results are shown in Table 2.


Table 2. Comparison of Abstract Generation Experiment Results.

| Model        | DUC-2004 |
|--------------|----------|
|              | ROUGE-1  | ROUGE-2  | ROUGE-L  |
| IR           | 11.06    | 1.67     | 9.67     |
| PREFIX       | 22.43    | 6.49     | 19.65    |
| Graph2Seq    | 29.30    | 8.39     | 24.46    |

From the experimental results, it can be seen that the summary generation algorithm based on the Graph2Seq model has improved the scores on the three evaluation indicators ROUGE-1, ROUGE-2, and ROUGE-L compared to the other two algorithms, which proves the method feasibility in improving accuracy.

4. Conclusion
The summary generation method based on the Graph2Seq model proposed in this paper incorporates keyword attention mechanism and output attention mechanism. Since graph neural network is used for training in the keyword extraction method, it is possible to improve the accuracy of keyword extraction while ensuring that the extracted keywords are closely related to the topic of logistics text. Through experimental comparison, it can be seen that the results based on the Graph2Seq model are superior to the existing abstract generation methods, which can be well used for the extraction of logistics news abstracts, effectively avoiding the wrong guidance of the sensational headline news, and improving reading efficiency.

Acknowledgments
This work was financially supported by Key Project of Beijing, China (Z191100001419001), Beijing Natural Science Foundation (4192042) and the Fundamental Research Funds for the Central Universities, New Teachers Program (2020RC14).

References
[1] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. Advances in NIPS, 2014.
[2] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
[3] Gehring J, Auli M, Grangier D, et al. A convolutional encoder model for neural machine translation[J]. arXiv preprint arXiv:1611.02344, 2016.
[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
[5] Rush A M, Harvard S, Chopra S, et al. A Neural Attention Model for Sentence Summarization[C]//ACLWeb. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2017.
[6] Niepert M, Ahmed M, Kutzkov K. Learning convolutional neural networks for graphs[C]//International conference on machine learning. 2016: 2014-2023.
[7] Yao L, Mao C, Luo Y. Graph convolutional networks for text classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 7370-7377.
[8] Xu K, Wu L, Wang Z, et al. Graph2seq: Graph to sequence learning with attention-based neural networks[J]. arXiv preprint arXiv:1804.00823, 2018.
[9] Mihalcea R, Tarau P. Textrank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 404-411.
[10] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2008, 20(1): 61-80.