Text Summarization Techniques Review

Zhouning Ma
zm11@illinois.edu

1. ABSTRACT

In recent years, the extractive summarization is still the most usage in production, because of the difficult technique of abstractive summarization and lack of readability and most extractive summarizations having better result. After Sequence to Sequence[6] was developed by google brain and proved having potential to performance better quality text summary than extractive summarization, more and more new solutions are implemented based on Sequence to Sequence architecture. And there was new solution published which mixed extractive summarization and abstractive summarization work together, it generated a better quality result than extractive summarization and abstractive summarization.

2. INTRODUCTION

Today's world is information explosion world and people receive huge amount of news, documents and text content. This amount of text requires to be effectively summarized for usage. It is impossible to do that all by human while automatic text summarization is an important technology to solve this problem.

Referred to Automatic Text Summarization [Juan and Torres 2014][1], there are couple reasons why automatic text summarization tools are important:

1. Summaries save your reading time;
2. It is easier to select the document
3. Improves the document index;
4. Save manpower;
5. Custom question-answering systems;
6. Some commercial services need short description when they can't support long text;

There are two methods of text summarization: Extractive Summarization and Abstraction Summarization. Most of the systems are implemented by extractive summarization method. In this paper I will introduce a new method: Mix Extractive Summarization and Abstraction Summarization together to implement a better quality solution.

3. SUMMARIZATION METHODS

3.1 Extractive Summarization

Extractive summarization is the task to choose the top rank sentences from the original text and group into a summary. Usually extractive summarization techniques don't change the sentences, but they will combine top K sentences as a short description.

There are couple tasks during the summarization process [Ani and Kathleen][2]

1. Use sentence classification or indicator representation to create intermediate representation;
2. Sentence Ranking will score to each sentence;
3. Pick the top K sentences to group together as a summary;

One technic which has been used in sentence ranking for a while is Graph-based [Salton ,Allan and Singhal][3] algorithms, such as Lex Rank, Page Rank algorithm. Graph-based ranking algorithms calculate similarly of the sentences to classify the sentences, the most common sentences get higher ranking [Rada][4].

There are a lot of document talking about extractive summarization. I list some papers talking about this topic [Mihalcea][5][Luhn][6][Mani][7]. In this paper I will focus on Abstractive Summarization and a new one Mix Summarization.

3.2 Abstractive Summarization

Abstractive summarization will generate the new sentences like what human reading the text and create the abstract. Abstractive summarization is full of challenge and the performance was not better than extractive summary in the past. After applying deep learning into abstractive summarization, the result has been improved, some of the result created by attractive summarization model is better than traditional extractive summarization and google brain's Sequence-to-Sequence is one of them.

3.2.1 Sequence-to-Sequence

Abstractive summarization condenses the text into a short description similar a human to create an abstract, it is more difficult to develop than Extractive summarization. The famous abstractive summarization is Google Brain team's algorithm called 'Sequence-to-Sequence' [Ilya, Oriol and Quoc][8].

Sequence to Sequence Learning use Deep Neural networks (DNNs) to achieve leaning tasks [Ilya, Oriol and Quoc] [5], [5] demonstrates how to use DNNs to translate English to French. Google team used the

sample ideal to apply on text summarization task. The algorithm creates two Long Short-Term Memory (LSTM) architecture to solve the DNNs doesn't support 'Memory' issue. One LSTM represented the input sequence, another one represented the output text. There is dependency between these two LSTMs.

Google Brain team uses this architecture to their Text Summarization by TensorFlow [Peter and Xin][9]. In their example: the input sentence is

> "*metro-goldwyn-mayer reported a third-quarter net loss of dlrs 16 million due mainly to the effect of accounting rules adopted this year*" [6]

and the output is

> "*mgm reports 16 million net loss on higher revenue*" [6]

Let's exam the result:

-"mgm" is the output of subject "metro-goldmyn-mayer", it looks like an abbreviation;
- The verb "reported" was changed to "reports" as output;
- It generated the conclusion "higher revenue" from the input;

There are another two examples (I didn't put here) which are good short descriptions and easy to read, it is hard for Extractive Summarization to create such type of result.

Google published the sql2seq source code on GitHub [Xin and Peter][10] and also published Experiment Result[7] too:

	Precision	Recall	F-score
ROUGE-1	0.50154	0.38272	0.42568
ROUGE-2	0.27565	0.20576	0.23126

Compared to 2004 Extractive Summarization algorithm ROUGE score, especial on ROUGE-2, the extractive summarization algorithm ROUGE-2 score is less than 0.09[Gunes¸ Erkan, Dragomir and Radev.][11], the Sequence-to-Sequence has a huge improvement.

3.3 Mix Extractive and Abstractive Summarization

As a new model for text summarization, it combines extractive and abstractive summarization call Mix Model, and the performance/quality is even better than abstractive summarization/extractive summarization, also it supports application/data scale.

3.3.1 Sequence-to-Sequence in 2018

Sequence to sequence learning is not only used for abstractive summarization but also used for extractive summarization. Some algorithms developed based on Sequence-to-Sequence are in

use today. There are two types of Recurrent Neural Network (RNN), Contextual Recurrent Neural Network and Non-Contextual Recurrent Neural Network.

Abstractive Contextual RNN(AC-RNN) and Extractive Contextual RNN(EC-RNN) are Contextual Recurrent Neural Network. AC-RNN passes a document context vector as input and a LSTM decoder and state-of-the-art techniques are used. EC-RNN use document-context-vector as an input as well, but decoding is using SoftMax, but the con is some sentences are always in result.

Abstractive Recurrent Neural Network (A-RNN), Extractive Recurrent Neural Network and Convolutional Recurrent Neural Network are Non-Contextual Recurrent Neural Network. These RNN didn't use document-context-vector when doing train.

3.3.2 Mix Model: Extractive and Abstractive work together to achieve better result

Seq2Sql models doesn't work well for long document (more than two lines of document). eBay developers created a solution combine Extractive and Abstractive summarization to work together. eBay implemented some Sequence to sequence with RNNs generate product snippet on its mobile app[Chandra, Gyanit and Nish][12] . eBay developers' solutions are:

1. Create context vectors from documents as preparation task;
2. Use large scale semi-supervised learning to run Extractive Summarization;
3. Run Extractive Summarization task by RNN and CNN-RNN;
4. Use step 1's context vectors to improve the Sequ2Seq learning;
5. Improve Abstractive Summarization by using the Extractive summarization result.

The training dataset is 768K words, human Extracted Snippets is 20K items/documents, 100K item descriptions, also sentences are tagged in total 100K items by looking into the Descriptions.

3.3.3 Evaluation the model

Using difference sentence long test cases to evaluation this model's performance (the dataset size is 15K for training and 5K for evaluation) [9], the results are:

- 1 sentence long and 3sentences long test result
  - EC-RNN is best on ROUGE-1: 0.39
  - E-RNN is best on ROUGE-2:0.37
  - TextRank ROUGE-1: 0.31
  - LexRank ROUGE-2: 0.25

- 5 sentences test result

-C-RNN is best on ROUGE-1:0.52
-V-RNN is best on ROUGE-2:0.61
-TextRank ROUGE-1:0.43, ROUGE-2:0.39

Anther 100K algorithmically labeled data test case, the ROUGE-1 and ROUGE-2 of EC-RNN and A-RNN are better than TextRank and LexRank.

4. CONCLUSIONS

Extractive summarization is the traditional method and still major use, after improved the quality of abstractive summarization by neural network being use, more and more abstractive summarization method will be used.

- In the past the drawback of abstractive summarization is the text product lack of readability, it is a big step of readability improvement by Sequence-to-Sequence solution, and i.e. eBay already used abstractive summarization into its production.
- The mixed model (Extraction summarization and Abstraction summarization together) integrated the strength of both and avoid the weakness, it shows a new method to implement text summarization, and it is not only using single method.
- Automatic evaluation methods still a big challenge, right now it still rely on human, and the dataset quality will affect the training result.
- Sequence-to- Sequence learning is pretty new and there is still a lot room to implement.

5. REFERENCES

[1] Juan-Manuel Torres-Moreno. Automatic Text Summarization

[2] Ani Nenkova and Kathleen Mckeown. 2012. A survey of text summarization techniques. In Mining Text Data. Springer.

[3] Salton, G.,Allan, J., Singhal, A.: Automatic text decomposition and structuring. Information Processing & Management.

[4] Rada Mihalcea: Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization.

[5] Mihalcea, R., tarau, P. TextRank: Bringing order into texts. In: Proceedings of EMNLP04 and the 2004 Conference on Empirical Methods in Natural Language Processing, 2004

[6] Luhn, H.P. The Automatic Creation of Literature Abstracts. IBM Journal of Research Development 2(2), 1958

[7] Mani, I. Automatic Text Summarization. John Benjamins Publishing Company, 2001

[8] Ilya Sutskever, Oriol Vinyals, Quoc V. Le. Sequence to Sequence Learning with Neural Networks arXiv.org

[9] Peter Liu and Xin Pan. Text summarization with TensorFlow Aug 24, 2016
https://ai.googleblog.com/2016/08/text-summarization-with-tensorflow.html

[10] Xin Pan, Peter Liu. Sequence-to-Sequence with Attention Model for Text Summarization.
https://github.com/tensorflow/models/tree/master/research/textsum

[11] Gunes¸ Erkan, Dragomir R. Radev. The University of Michigan at DUC 2004. Page 7 (Table 6: Results for Task 2)

[12] Chandra Khatri, Gyanit Singh, Nish Parikh. Abstractive and Extractive Text Summarization using Document Context Vector and Recurrent Neural Networks. Page 1.