Multi-Document Summarization Applied to Drug Reviews

Darin Maduar
Faculty of Computer Science, Higher School of Economics, Moscow, Russia
drmaduar@edu.hse.ru

***Abstract*—In recent years, the number of reviews and product descriptions has been increasing rapidly. Usually, it is difficult for the average user to process even the most important of these reviews. The most valuable solution for this problem is a text summarization. The purpose of this article was a review the existing multi-document summarization algorithms. In addition, we will discuss the possibility of applying these algorithms to the drug review summarization task. We want to introduce an algorithm that generates a single synthetic description containing the most important information about the drug through multiple drug reviews.**

***Keywords*—Natural Language Processing, Text Mining, Text Summarization, Multi-Document Summarization.**

Introduction

Nowadays, any user of online shops can leave a product review. There can be many such reviews, some too long, others too uninformative. In addition, reviews often contain similar information about a product. It is difficult for a person to process so much information. However, this information could be very useful. If we are talking about drug review summarization, user opinion about the product can help drug pharmaceutical manufacturers analyze efficacy or collect statistics about side effects. It is well known that all drugs are tested with a series of clinical trials before they go out to market. But there are many relatively recent studies that point to the importance of postmarketing drug surveillance[1], [2].

Automatic text processing is a research field that is currently extremely active. One important task in this field is *automatic summarization*, which consists of reducing the size of a text while preserving its information content [3], [4]. A *summarizer* is a system that produces a condensed representation of its input's for user consumption [5] [6]. As regards reviews, they generally contain some special *aspects* (i.e., disease, treatment period, dosage, efficiency...). It would be useful to collect information for each aspect. So, we want to focus on *multi-aspects review summarization* [7] task.

We will use Drug Review Dataset data set. This is data obtained by scanning sites with pharmaceutical reviews. It contains the following information:

- name of drug;
- name of condition;
- patient review;
- 10 star patient rating;
- date of review entry;
- number of users who found review useful.

This data set became publicly available after the end of the ”UCI ML Drug Review dataset” Kaggle competition.

Literature review

The problem of text summarization has been studied for a long time, such as (Luhn, 1958 [3]; Kupiec, 1995 [8]). However, applying traditional summarization methods directly on drug reviews doesn't give acceptable results.There are several reasons for this. First, the number of product reviews is often much larger than the data. Secondly, product review sentences are usually conversations and contain a lot of uninformative words. Directly retrieved resumes may contain a large amount of undesired information.

A number of researchers have studied the task of review summarization. (Ganesan et al., 2010 [9]) proposed a graph-based method for generating ultra concise opinion summaries of products. They used predefined rules for finding valid sub-paths in the graph and converted those sub-paths into sentences. Since the sentence generation was rule-based, their method didn't provide a well-formed grammatical summary. (Gerani et al., 2014 [10]) generated product review summaries by using discourse structure. After simplifying the discourse graph, they used a template-based NLG framework to generate natural language summaries. Their summary produced a statistical overview of the product but lacked detailed information. (Ganesan et al., 2012 [9]) proposed some heuristic rules to generate phrases, they used a modified mutual information function and an n-gram language model to ensure the representativeness and readability of the phrases. However, their method didn't consider the descriptiveness of the phrases.

Methods

To begin with, we plan to implement the existing method of summarizing reviews described by (Yu et al., 2016 [11]). There proposed a phrase-based summarization algorithm for the task of *product review summarization*. The proposed phrase selection scheme can fully utilize the characteristics of review sentences and capture the main information. And we'll try to apply this algorithm to our data set. Analyzing the results, we will try to adjust the algorithm to the specifics of our problem.

Results anticipated

We look forward to receiving a framework that will generate good aspect-based review summarization for each medicine.

The evaluation of review summarization is a very challenging task. On one hand, to the best of our knowledge, there is no dataset of drug reviews with human written summaries. On the other hand, since the amount of drug reviews is often large, it is quite difficult to generate human written summaries. We'll try to generate the summaries by our own system and evaluate with a state-of-the-art extractive baseline, a state-of-the-art abstractive baseline and a simple BasicSum algorithm (Yu et al., 2016 [11]). The details of the baselines are described as follows:

1) **LexRank**: LexRank (Erkan and Radev, 2004 [12]) is a graph-based extractive summarization method which computes sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. In the experiment, first we cluster sentences by their aspects. Then for each sentence cluster, LexRank is performed for summary generation. The final summary is generated by putting summaries of different aspects together in the same aspect order of ReviewSum.
2) **Opinosis**: Opinosis (Ganesan et al., 2010 [9]) is a novel graph-based summarization method which generates concise abstractive summaries of highly redundant opinions. In the experiment, for each aspect, we build an Opinosis graph and get the top candidate summaries. The final summary is generated by putting summaries of different aspects together in the same aspect order of ReviewSum.
3) **BasicSum**: BasicSum is a simplified version of our summarization method. Instead of popularity and specificity, TF-IDF score is used in the objective function. The objective function of BasicSum can be denoted as:

$$F(x_1, \cdots, x_n) = \sum_i tfidf(p_i) \cdot x_i$$

where $tfidf(p_i)$ is the TF-IDF score of phrase p_i , and x_i is a binary value representing whether phrase pi is selected in the final summary or not.

Conclusion

Within this preliminary work, we studied the application of machine learning based sentiment analysis of patient generated drug reviews.We made an overview of algorithms that solve similar problems and identified those that would be suitable for solving our problem. In addition, we investigated the importance of the task for postmarketing drug surveillance.

References

[1] John Mcneil, Loretta Piccenna, Kathlyn Ronaldson, and Lisa Demos. The value of patient-centred registries in phase iv drug surveillance. *Pharmaceutical Medicine*, 24:281–288, 10 2010.

[2] E. O'Brian Smith Rose Lee Bell. Clinical trials in post-marketing surveillance of drugs. pages 61–68, 3 1982.

[3] H. P. Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165, April 1958.

[4] K. Spärck Jones. Automatic summarising: the state of the art. *Information Processing and Management*, 43(6):1449–1481, 2007.

[5] Udo Hahn and Inderjeet Mani. The challenges of automatic summarization. *IEEE Computer*, 33(11):29–36, 2000.

[6] Joel Neto, Alex Freitas, and Celso Kaestner. Automatic text summarization using a machine learning approach. volume 2507, pages 205–215, 11 2002.

[7] Ryosuke Tadano, Kazutaka Shimada, and Tsutomu Endo. Multi-aspects review summarization based on identification of important opinions and their similarity. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 685–692, Tohoku University, Sendai, Japan, November 2010. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

[8] Julian Kupiec, Jan O. Pedersen, and Francine Chen. A trainable document summarizer. In *SIGIR '95*, 1995.

[9] Jessica R. Grubb, Edgar Turner Overton, Rachel Presti, and Nur F. Önen. Reply to Ganesan et al. *The Journal of Infectious Diseases*, 205(3):518–519, 11 2011.

[10] Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613, Doha, Qatar, October 2014. Association for Computational Linguistics.

[11] Naitong Yu, Minlie Huang, Yuanyuan Shi, and Xiaoyan Zhu. Product review summarization by exploiting phrase properties. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1113–1124, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[12] G. Erkan and D. R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.