Opinion generation using abstractive Text Summarization techniques

Sumedh Sankhe, Soumitra Mishra and Rajath Kashyap

1 Introduction

We live in a world which is data driven, where large amounts of data is generated every second from every digital process. With this exponential growth of data, abstracting or summarizing the generated data becomes a necessity for its effective consumption. With this data overload, it is very important for a user to quickly and effectively understand and use this data for their needs. This is where automatic text summarization plays a crucial role in solving the problem.

Text summarization can be defined as the process of generating a concise, sensible text from large amount data, while still preserving the important information and meaning from the original text. In our project we try to tackle the problem of generating a generic opinion of a product or service by using customer reviews on these products or services. We can find a large amount of customer reviews on almost every e-commerce platforms like Amazon, Zomato, Yelp to name a few. We can find the overall star rating of any product or service but to get a general review of the product, the user has to read through hundreds of reviews. So in this project we will generate a concise generic opinion of the product by reading and understanding the reviews, extract important information from them and condensing these large number of reviews to generate a small summary portraying the generic idea of the reviews.

2 Related Work

- In 2012 Ganesan et.al [2] proposed an Unsupervised with some heuristic algorithm to generate ultra concise summaries of opinions, this method treats the problem as an optimization problem where and measures the representation base on a modified mutual information function and an n-gram based langage model.

- In 2008 Branavan et.al [3] proposed to generate concise summaries of opinions by implementing a heirarchial bayesian model reusing the pron and cons phrases to train their keyphrase extraction mode. The approach focused on simultaneously finding hidden paraphrase structures, model of document texts and the underlying semantic properties that link the two which allowed the use of unannotated documents.

- Extractive or Natural Language Generation (NLG) have been two different summarization techniques used Carenini [4] compared both the methods for opinion summarization by defining a novel measure of controversiality.

- Latent Semantic Analysis (LSA) identifies semantically important sentences for summary creation whereas traditional IR methods measure sentences relevance, both these methods strive to select sentences that are highly ranked and different from each other, proposed by Gong and Liu [5]

- SUMMONs uses conceptual summarization rather than linguistic summarization in order to generate natural language summaries from multiple online sources ranging from live news streams to the CIA World Factbook, and past newspaper archives, proposed by Radev and McKeown [7]

3 Overview

Text summarization can be broadly classified into two types[1]: Extractive and Abstractive. In Extractive summarization the summary is generated by extracting the most important and useful sentences from the given document and using the same sentences in the summary. In an Abstractive summarizer, after extracting the important information from the document we paraphrase sentences to generate a summary which makes it more human like summarizing. In this project we make use of the Abstractive concept, to show, typically a review of the product containing a couple of phrases that describe the opinion rather than showing a sentence picked from the review.

During the initial phase of the project, we would like to explore on the various text summarization techniques, the features to be leveraged out of the given

text and select a combination of techniques to build the model.

We will start of by cleaning and structuring the user reviews in format that fits our model, ideally one document per product. Every document will contain all the reviews associated with that particular product. Second, we would like to define certain desirable characteristics that are exhibited by our words namely,

- Representativeness - It should accurately present the critical information like major opinions, complaint or praise of the product under consideration.
- Readability - Summary should be well organized, grammatical and should make a sense to the reader.
- Compact - It should be concise approximately 8-10 words.

Further as an addendum to the initial research carried out by Ganesan et. al [2] we will classify the generated summary into pros and cons about a given product. A [12]Naive Bayes classifier will be appropriated for this case.

4 Dataset and Evaluations

We will use the data from CNET as the primary dataset, which is based on user review on CNET. These reviews lie in various categories, mainly electronics, like mobile phones, televisions, laptops etc. The other datasets that will be used include the Amazon reviews, Yelp reviews, and the Zomato review datasets which will be our secondary datasets. We will partition our primary dataset into testing, heldout testing datasets. Recall-Oriented Understudy for Gisting Evaluation[11] (ROUGE) and Bilingual Evaluation Understudy (BLEU) are the two most commonly used metrics to score text summarization we will be using ROUGE to measure the performance of our model

References

[1] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text Summarization Techniques: A Brief Survey. In Proceedings of arXiv, USA, July 2017,

[2] Kavita Ganesan, ChengXiang Zhai, and Evelyne Viegas. 2012. Micropinion Generation: An Unsupervised Approach to Generating Ultra-concise Summaries of Opinions. In Proceedings of the 21st International Conference on World Wide Web (WWW '12). ACM, New York, NY, USA, 869–878.

[3] S. R. K. Branavan, H. Chen, J. Eisenstein, and R. Barzilay. Learning document-level semantic properties from free-text annotations. In In Proceedings of ACL, pages 263–271, 2008,

[4] Giuseppe Carenini , Jackie Chi Kit Cheung, Extractive vs. NLG-based abstractive summarization of evaluative text: the effect of corpus controversiality, Proceedings of the Fifth International Natural Language Generation Conference, June 12-14, 2008, Salt Fork, Ohio,

[5] Y. Gong, X. Liu: Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States 2001,

[6] K. McKeown, J. Robin, K. Kukich: Generating Concise Natural Language Summaries. Information Processing and Management: an International Journal, Volume 31, Issue 5, 1995,

[7] R. Radev, S. Teufel, H. Saggion, W. Lam, J. Blitzer, H. Qi, A. Celebi, D. Liu, E. Drabek: Evaluation Challenges in Large-scale Document Summarization. Proceeding of the 41th annual meeting of the Association for Computational Linguistics, Sapporo, Japan 2003,

[8] Landauer TK , Foltz PW , Laham D. An introduction to Latent Semantic Analysis . Discourse Processes 1998; 25: 259-284,

[9] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: a graphbased approach to abstractive summarization of highly redundant opinions. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 340–348.

[10] Regina Barzilay and Lillian Lee 2003. Learning to Paraphrase: An Unsupervised approach using multiple sequence alignment. In Proceedings of HLT/NAACL 2003. Department of Computer Science Cornell University Ithaca, NY: 14853-7501.

[11] Josef Steinberger, Karel Jezek 2007. Evaluation Measures for Text Summarization. Department of Computer Science and Engineering University of West Bohemia in Pilsen, 306 14 Plzen, Czech Republic.

[12] Bharath Sriram, Murat Demirbas, Dave Fuhry text classification in twitter to improve information filtering Ohio State University, Columbus, OH, USA Published 2010 in SIGIR