Develop a model for abstractive summarizaiton on Amazon reviews

Xinyue Ke, Lu Pan, Chang Liu
Department of Computer Science
University College London
ucabxke;uczllpa;uczlcl8@ucl.ac.uk

Abstract

In recent years, the use of e-commerce has been grown rapidly and then online shopping has become the essential part of most people in life. People may write the reviews after shopping online. This character of online shopping could not only help merchants to increase sales but also help customers to choose the right products from a large number of products. However, there are also disadvantages to customers and merchants. For customers, it is sometimes difficult for them to choose the right information from reviews that they really want to know. This is because for the popular products, the number of reviews could be thousands or even tens of thousand. The disadvantage to merchants is that it is difficult for merchants to manage and track the reviews from customers. In this paper, we aim to develop a model to do the abstractive text summarization of customer reviews from the amazon food products. We do not focus on the customers opinions (positive or negative) since we want to capture the features of a product and mine the reason behind the customers sentiment. In this paper, we propose three model frameworks which are the extension of attentive recurrent neural network. By developing the model, we will make it easier for customers to make their purchasing decisions. The effectiveness of our experiment could be demonstrated by comparing the summarization result with the feed-forward neural network model which is the baseline of our experiment.

1 Introduction

With the widespread use of e-commerce platform, buying commodities online has become the main way of shopping. More and more people and modern merchants has taken part in the conventional commercial activities. At the same time almost all online shopping websites allow and encourage people to leave reviews after buying commodities. Thus, the textual data expand dramatically in the modern e-commerce platform. Good news for customers is that they could know the features of a product easier from previous guests in contrast to the past times without online shopping. However, e-commerce has developed to present and been mature. Consequently, the number of reviews for popular products become extremely huge. Furthermore, the length of some reviews are very long and only a few sentences contain the valid information. For customers, various and huge amount of products with so many reviews may lead to the difficulties in making purchasing decisions. Customers may spend a lot time browsing comments from other customers, but they finally get the biased view. Merchants also have difficulties in managing their own business and improving the quality of services they provide.

Many projects now focus on determining the sentiment polarity by exploring the methods of opinion mining, but this is not enough for customers to make purchasing decision. For example, there is a type of food on amazon which is organic but with poor taste. A customer, who pay much attention to organic food but does not care about the taste, could leave reviews with positive polarity. Only extracting the sentiment polarity of the food may result in another customer who focus on taste to buy the food. Therefore, the accurate summary of features of a product from reviews could satisfy the unique requirement of customers to the product. In addition, it is also helpful for merchants to adjust their own goods and meet the requirement of customers.

In this paper, we face above challenge by generating the feature based summaries of multi-product reviews which were collected from Amazon. Text summarization (Singhal and Bhattacharya, 2017), the active field in Natural Language Processing, is the process of automatically

generating the summarization with major points of a given input document. There are two types of evaluated systems for text summarization: Extractive systems and abstractive systems. Extractive systems (Singhal and Bhattacharya, 2017) create the summaries by copying a subset of words which could reflect the main topic of the given text, whereas abstractive systems generate summaries by using new phrases and words that are not in the original text. Much work has been done in the area of sentiment analysis and extractive summarization , but work on the abstractive summarization of reviews is less available.

In this paper, we focus on the abstractive summarization of multi-product reviews from Amazon. We propose three models which are the extension of attentive recurrent neural network. In the evaluation part, we consider three quantitative evaluation mechanisms ROUGE-N, ROUGE-L and METEOR. We regard feed-forward neural network model as our baseline.

2 Literature Review

2.1 Related Work

As customers reviews attract more attentions than usual in recent years, there had been lots of work about review summarization. In this section, we would analyze several methods such as genre and sentiment classification, extractive summarization and abstractive summarization, and we would especially focus on the last one. Sentiment analysis is the basic task of opinion mining. By using a dataset with a label(e.g. positive or negative) on each review, we could build a sentiment classifier. This kind of model would perform well and have a high accuracy on the test dataset. However, there is an obvious shortage that the information we get from sentiment analysis is not enough, because the potential customers might want to know more about the features of the products. To solve this problem, we introduce automatic summarization which could generate a shorter version of given sentences while preserving its meaning to the utmost extend. There were two main ways of text summarization, extraction and abstraction. The extractive system was to select some part of the source text based on the linguistic and statistical importance and stitch them together to produce a sensible and shorter version. In contrast, abstractive syetem could use novel phrase that might not be contained in the source text. Extractive system is a easier task and could ensure a baseline level of syntactic and lexical accuracy. On the other hand, understanding of common sense and summarize in a more efficient and general way could only possible on the abstractive system. Our work is related to Rush (2015) where the author implemented a Neural Attention Model with different encoders. This model is related to the standard feed-forward neural network. However, on the basis of their work we propose the extension of recurrent neural network with LSTM and attentive-based methods.

2.2 Sentiment Classification

Hearst (1992) introduced a direction-based classifier which could output three directionality (e.g. like, neutral, dislike) for the whole source text. Huettner and Subasic (2000) created a small set of semantic categories and constructed a affect lexicon with each word assigned to two scores corresponding to each categories. These two scores represent for centrality (degree of relatedness) and intensity (strength of the emotion) respectively and they could be used as a quantitative measurement of sentiment. Das and Chen (2001) developed a method for extracting the information of relation between the stock price and the views of small inventors. By combining 5 different classifiers, message were classified into 3 types (bullish, bearish and neutral). Turney (2002) classified reviews into recommendation and not-recommendation based on average semantic orientation of adjectives and adverbs by unsupervised learning algorithm. Semantic orientation were calculated as common information between the given phrase and excellent minus the common information between the given phrase and poor, and a view was classified as recommended if the score was positive and verse vise. In the same year, Pang *et al.* (2002) treated this problem as topic classification with two topics positive and negative, since topical categorization had been studied more and had a high performance. However, as it has been mentioned in the introduction, simple sentiment classification could only preserve the polarity of each review. What customers really need to know is the characteristics of products and the reason of the sentiment. Thus, in order to overcome these disadvantage, we focus on the text summarization. Then we would analyse the work related to the extractive summarization.

2.3 Extractive Summarization

In 1958, Luhn published a paper in which he described the possible simplest extractive system. The fact is that the writer may repeat the words which is related to the topic of his article. This means that emphasis on the frequency of words in an article is reasonable. Thus, Luhn suggested that the words with the highest number of occurrence could reflect the main topic of an article. We actually know now this method does not work most of the time. For example, an entertainment news for sport stars may be recognized as the sports news since the words related to the topic of sports could appear with the highest frequency. In our case, most of the reviews from customers are short and then keywords may appear less times. In addition, this method could not extract the opinion of customers. Thus, this method could not work well.

In order to improve the efficiency of the task of keyword extraction, other methods based on lexical and syntactic features were explored. Some keyphrase extraction methods appeared. Keyphrase extraction is referred to the automatic selection of a set of words that could represent the topic of the given document. In addition, the main idea of keyphrase extraction is to transform the process of keywords selection into the binary classification problem. More specifically, each word in the given document whether a keyword or not. The most important factor to classify a keyword candidate is that the frequency and location of a word in the document.The section below describes several keyphrase extraction methods.

In 1999, Turney published a paper in which he firstly suggested GenEx (Genitor plus Extraction) system, a supervised learning algorithm, which is the keyphrase extraction method. According to the definition provided by Turney, GenEx is the combination of Genitor genetic algorithm and the Extractor keyphrase extraction algorithm. Having defines the what is meant by GenEx, we will now describe the process of keyphrase extraction by the GenEx system. In the training process, both Extractor and Genitor genetic algorithm are needed. There are 12 parameters in the Extractor. These parameters in Extractor could determine the process of keyphrase extraction from the given input document. Genitor genetic algorithm could be used to adjust the parameters in Extractor so that we could maximize the performance of the system on input training document. Once we find the best parameter values, the training process is completed and the Genitor can be discarded. In other words, Genitor assists the Extractor to find the best parameter values in training process. After the training process, Extractor is used to do the keyphrase extraction. However, the training process of GenEx is high computational complexity.

In order to solve the problem of computational complexity, a different keyphrase extraction method which is based on Naive Bayes Learning scheme was reported by Frank in 1999. This method is known as Kea keyphrase extraction algorithm. Frank obtained an improved result by applying Kea on the same dataset used by Turney in 1999, but training process is much quicker. Kea algorithm could be divided into two stages: training and extraction. In the training process, after cleaning and organising the input document Kea algorithm firstly generate the candidate phrases according to some rules. Then the algorithm calculate two features for each candidate phrase. The two features are $TF \times IDF$ score of a candidate phrase and the distance into the given document of the phrases first occurrence. The term $TF \times IDF$ is defined to measure the how specific a phrase P is in the given document D:

$$
\begin{aligned}
& TF \times IDF \\
=& P(\text{phrase in D is P}) \times [-\log P(\text{P in a document})] \\
& \text{where} \\
& P(\text{phrase in D is P}) = \frac{\text{the number of phrase P in D}}{\text{size of D}} \\
& \text{and} \\
& P(\text{P in a document}) \\
=& \frac{\text{documents containing P in the training corpus}}{\text{total number of document in training corpus}}
\end{aligned}
$$

The second feature distance is calculated as the number of words that appear the phrases first occurrence, divided by the number of words in the document. After obtaining the two real features values, kea algorithm applies the Bayes formula to calculate the probability that a phrase is a keyphrase given that the discretized $TF \times IDF$ and discretized distance D under the assumption that $TF \times IDF$ and distance D are independent. Kea algorithm built a Naive Bayes model from the training documents for which keyphrases are known by the above procedure. We finally apply the resulting model to a new document from which keyphrases are required to be extracted.

The two algorithms GenEx and Kea are extremely important in the field of keyword extraction since they are the foundation of the next algorithms. Furthermore, they have become the baseline when other methods evaluate their own performance.

As far as keyphrase extraction methods are concerned, turning now to a graph based keyword and sentence extraction method. TextRank (Mihalcea and Tarau, 2004) was formally introduced by Mihalcea and Tarau in 2004. it is an unsupervised algorithm and belongs to the extractive system. The most applications of TextRank are keyword extraction and sentence extraction and then use them to form the summarization. As a graph based ranking model, the basic idea of TextRank is voting. In order to find keywords and the most relevant sentences in the given document, a graph containing vertices and edges is constructed. For keyword extraction, this algorithm regards words as vertices. Edges are added between two words according to the co-occurrence relation. In other words, two words are connected if these two words occur at the same time within a window of maximum N words ($2 \leq N \leq 10$). For sentence extraction, the graph vertices are sentences. Two sentences could be connected if there is certain content overlap. After ranking words and sentences in the graph, TextRank could find the most representative words and sentences for the given text.

Algorithm mentioned above belong to extractive system which do not involve the too much sentiment analysis. In contrast to normal articles, the language of customers reviews is more special since it could contain some specific elements which are not too much in normal articles. These elements could be irony and sarcasm. Customers also could express the identical thing in completely different way. Furthermore, most of the reviews from customers are short and then keywords may appear less times. Thus, these methods could not work well. We aim to develop the abstractive system to summarize the multi-product reviews from Amazon.

2.4 Abstractive Summarization

In this section, we would discuss the paper published by Rush, Chopra and Weston in 2015 in more depth. This paper proposed an abstractive summarization system, a Neural Attention Model, which is a fully data-driven approach.

We define the scoring function as

$$s(x, y) \approx \sum_{i=0}^{N-1} g(y_{i+1}, x, y_c)$$

and our goal is to maximize the scoring function.

In particular, if we consider the scoring function as conditional log-probability, our goal is to find the sequence that maximize the following function

$$\sum_{i=0}^{N-1} \log p(y_{i+1}|x, y_c; \theta)$$

So our task was to model the local conditional distribution

$$p(y_{i+1}|x, y_c; \theta)$$

We would describe the architecture of the Neural Attention Model in detail. The network in this model is a combination of a neural probabilistic language model and an encoder.

Neural Probabilistic Language Model: The language model is constructed adapted from the standard feed-forward neural model. It could be used to calculate the probability of a word in a sentence, given the input sentence and previous C words. Rush proposed this model as following:

$$\begin{aligned} p(y_{i+1}|y_c, x; \theta) &= \text{softmax}(Vh + Wenc(x, y_c)) \\ \tilde{y}_c &= [Ey_{i-C+1}, \ldots, Ey_i] \\ h &= \tanh(U\tilde{y}_c) \end{aligned}$$

Encoders: The encoder in the network could be regarded as a conditional summarization model. We could consider different encoders in the network. Turning now to describe these encoders that Rush listed in the paper.

Bag-of-Words Encoder: Bag-of-Words could represent a text by describing the occurrence of words within the text. This encoder do not focus on the order and relationship between words. words with the high frequency in the given text could get heavily weighted. Thus, the weakness of this encoder is that the relationship between neighbour words is inherently limited.

Convolutional Encoder: This encoder could solve the issues existing in the Bag-of-Words Encoder by involving the local interactions between words. In addition, this encoder does not require the information from previous words. However, convolutional encoder only generates one representative word for the whole input sentence.

Attention-Based Encoder: Essentially, this encoder uses a simple model similar to the bag-of-words encoder. We paid different attention to different context by replacing the uniform distribution in bag-of-words with a learned weight matrix parameter. The words $x_{i-Q}, \ldots x_{i+Q}$ were highly weighted by the encoder if the current context aligned well with position i where Q was a smoothing window.

Training: Having discussed the architecture of Neural Attention Model, turning now to the training process. When training our model, we need to minimize the negative log-likelihood over the training set consisting of all input-summary pairs. This process could be done by using mini-batch stochastic gradient descent.

After the training step, we could get all the conditional probability distribution of first N words. Since Viterbi decoding requires $O(NV^C)$ time which could be extremely large when V is large, then we could choose a decoder between finding a exact solution and strictly greedy, named beam-search, required $O(KNV)$ time where K is the beam size. In each iteration, we only need to compute the conditional probability for each word on the condition of the K hypotheses and keep the highest first K words where these K hypotheses are computed and keep in the previous step.

3 Project Proposal

3.1 Objectives

The focus of this work is to understand the frameworks applied in abstractive summarization in natural language processing and to develop a recurrent neural network models that can generate improved-quality abstractive summaries for reviews on Amazon.

3.2 Evaluation

For the models we will investigate, quantitative evaluation mechanism ROUGE-N, ROUGE-L (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) to evaluate the methods and qualitative analysis by manual inspection are both involved. For the quantitative part, ROUGE-N is an n-gram recall between system summaries produced by machine and a set of reference summaries labelled by human. In this research, ROUGE-1, ROUGE-2 would be used. And ROUGE-L requires sequence level matches rather than consecutive ones and automatically identifies the maximum length of co-occurring n-grams in sequence. In the meanwhile, METEOR is used to test the standard exact word matching, stemming and synonymy matching. However, the quantitative analysis seems not fully reliable and is insufficient. Thus visual inspection of results is necessary to examine the quality of summarization.

The feed-forward neural network model is used as a baseline in this research. The results obtained from our models will have a comparison with this baseline.

3.3 Data acquisition

We downloaded Amazon multi-product reviews dataset from Stanford Network Analysis Project (SNAP).This dataset contains thirty five million samples. Reviews include product and user information, ratings, and a plaintext review. For each sample, we select a summary of this review and a complete review text. The invalid data that the reference summary inside did not summarize the content appropriately would be removed in the prepossessing stage.

3.4 Existing frameworks and models in this research

Rush *et al.*(2015) developed a model using an attention-based encoder and a beam search decoder. The beam search maintains the whole vocabulary while selecting top k hypotheses at each stage of the summarization. The scores was not as good as expected. Then a model with improved results was proposed in 2016 by Chopra et al. RNN is a kind of neural networks with loops in them, which is different from feedforward neural networks, as it can store memory by using their internal state. LSTM is a variant of basic RNN layer, which is capable of handling long sequences. LSTM is formed by a cell, an input gate, an output gate and a forget gate. The cell stores values and the gates control the flow of information pass through the cell. This model used convolutional attention-based conditional recurrent neural network in the encoder part. Firstly, the full embeddings was obtained by calculating the summation of learnable position embedding and word embedding. Secondly, a 1d kernel, a learnable weight matrix, was designed to convolve over the full embeddings of consecutive words, further generating aggregate embedding vector. Then after applying dot product attention, the softmax function was used to normalise

it. Thus, in the encoder part, the output in time t was calculated by the sentences and the previous hidden states from the time t-1 decoder layer. In the decoder part, RNN and LSTM was used and the hidden state was obtained then sent to the encoder part for the next operation. This model is optimized by mini-batch stochastic gradient descent. These studies give us insight of using attention-based encoder and RNN to build a summarization model. Our research is an extension of these works and focuses on Amazon reviews.

Several changes towards the model focusing on Amazon reviews can be applied. These tricks may improve the quality of summarization. For example, there are different mechanisms of attention system. For the attentive algorithm, besides dot product, general, concat and perceptron are also appropriate choices. And local attentive method may perform better than the global one (Luong, 2015). Also, there are works which present a general attention-based CNN for modeling a pair of sentences (Yin, 2016). In ABCNN-1, attention part is completed before convolution starts, by using attention matrix to compute the attention feature map of corresponding sentence pairs. ABCNN-2 processes attention part while pooling. ABCNN-3 combines ABCNN-1 and ABCNN-2 by stacking them. This paper offers us the insight about diverse ways to apply attention in CNN. In our work, we can use two attention layers, one for sentences and one for words.

The architecture of RNN system can be improved as well. We can use Bidirectional RNNs, or BiLSTM in the encoder part. This is based on the consideration that the output should not only be related to the former information, it is also influenced by the latter counterparts. BiRNN concatenates two hidden layers, which are in forward and backward directions, to generate a final output. With this form of generative deep learning, the output layer can have information from past and future states simultaneously.

To sum up, our work will be a implementation an extension of attentive recurrent neural network proposed by Chopra *et al.* (2016). A more sophisticated attention mechanism and Bidirectional RNNs will be implemented in models this research to seek an improvement. Three model frameworks we will implement apart from the baseline are:
Model 1 attention-based CNN encoder and a RNN decoder
Model 2 an improved attention-based CNN encoder and a RNN decoder
Model 3 an attention-based BiRNN encoder and a RNN decoder
where RNN can be LSTM.

This design is a trade-off of accuracy and speed considered training RNN is quite slow compared to training CNN. But we still want to investigate how BiRNN can affect the quality of summarization. Thus more sophisticated models are not in our consideration. There are resources suggested other possible methods to improve quality of summarization. For example, involving reinforcement learning in our models may generate a better result. Some studies (Chen, 2018) showed using ROUGE to optimize the output is more effective than minimizing negative log-likelihood. But whether these methods will be implemented will depend on our project progress. Besides, Dropout can be used in CNN layer to prevent overfitting.

3.5 Existing open-source project

As stated, our work is an extension of attentive recurrent neural network proposed by Chopra *et al.* (2016). This work has an open-source at: https://github.com/facebookarchive/NAMAS. Our models are built upon it. Besides, the open source of the baseline model can be built based on https://github.com/facebookarchive/NAMAS.

3.6 Project steps

The steps for measuring progress of this project is laid out as follows:

1. Prepossessing Amazon review set to obtain high-quality data.
2. Implement and evaluate the baseline model based on the open-source project.
3. Use a small subset of data and try parallel ideas about Model 1, Model 2 and Model 3 designed above. i.e. whether we should use basic RNN, LSTM or GRU and which attention system should we choose.
4. Evaluate them and pick state of the art models among these trials.
5. Reproduce results with models selected and apply model-specific optimizations i.e. hyperparameter choice.
6. Evaluate the results and find the hypermeters in models with best scores.
7. Compare these results to the baseline model. Analyse the factors that can affect the result.
8. Write a paper to report this project.

4 Feasibility Study

4.1 Data and descriptive statistics

We obtained a dataset consisting of reviews on Amazon from SNAP. The dataset spanned over 18 years from Jun. 1995 to Mar. 2013, including about 35 million reviews. Reviews in the dataset are from some kinds of products such as, books, electronics, movies and TV, toys, sports, music, etc. Entries for each item contain the following: products, price, user information, ratings, time of reviews, original review and review summary. In the raw dataset, there are 34,686,770 reviews for 2,441,053 products which were written by 6,643,669 customers. 56,772 customers have left more than 50 reviews. In addition, the median of the reviews in the raw dataset is 82.

The above section have given the background of the raw dataset. The next part will describe the process of generating the dataset we will use in our research. We totally have 24 topics with different number of reviews. Our aim is to generate the feature based summaries of customers reviews. After looking at these topics, we give up some topics, such as Beauty and Clothing, Shoes and Jewelry, since summaries of reviews in these topics are short and tend to contain more sentiment to products. These topics are more suitable to do the sentiment analysis. According to our research aim, we finally picked the following 3 topics: Electronics, CDs and Vinyl and Movies and TV. The reason for us to choose the 3 topics is that summaries of reviews in the three topics contain more information of products and less sentiment to products than that in other topics. The characters of summaries in the 3 topics are consistent with our aim.

Having finished the selection of topics, turning now to the data we will use. The three topics we choose totally contain 4,484,313 reviews which are large. Then it is impossible for us to put all of them into our model. Furthermore, some summaries of reviews in the three topics are still not suitable for us to do the feature based summarization. Thus, we dealt with the whole dataset as following. We mixed all data from the three topics. Then we deleted the data for which the number of words in summaries is below 3 since these data could only contain the sentiment to the products. After deleting the data that meets the above requirement, we randomly selected around 300,000 data from the rest data. We plan to utilize around 240,000 reviews for training, around 15,000 for testing and around 45,000 for validation.

We expect that our model could generalise to the multi-product reviews summarization. In other words, we expect that our model could work on the products that do not appear in the training set. Thus, in order to test the above point, we added the additional data to the test dataset. We chose one more topic Toys and Games which contains 167,597 data. Then we deleted the data for which the number of words in summary is below 3. After deleting the data, we randomly selected around 5000 data from the rest data. Finally, we added the 5000 data to the test dataset. Our test dataset now contains around 20,000 data. Overall,we use around 240,000 for training, around 20,000 for testing and around 45,000 for validation.

The above part has demonstrated the process of generating the dataset we will use in our research, we now give the descriptive statistics of it. It would be helpful to know our dataset before we started to train and test our model. In our dataset, the mean of the length of reviews is 102 words, the median is 82, and the interquartile range of the length is 76 to 91, which means that the length of reviews are positive skewed. The average length of the summary is 8, the median was 7, and the interquartile range is 5 to 9, which means that the length of the summaries were positive skewed.

We have done much preparation for the data which is from the three selected topics. We firstly delete question marks, semicolon, punctuation and colon. Then we convert all words to lowercase.

4.2 Qualitative Examples

In this part, we will give several qualitative examples to demonstrate the feasibility of our research. In order to do the qualitative analysis, we picked three reviews from our dataset which are paired with the reference summary. The reference summary refers to the original summary of the selected review text. Then we let three people analyze the original review text and give us three summaries for each review text. Three qualitative examples are listed as following:

4.2.1 Qualitative Example 1

Review: Henry Winkler is very good in this twist on the classic story. Not a conventional remake, but a version of the story set in early America. Give it a try.

Reference Summary: It's an enjoyable twist on the classic story.

Possible Summary 1: Good classic story in America without a conventional remake.

Possible Summary 2: It worths a try since the good twist in an early American story.

Possible Summary 3: It's a good twist on a classic story.

Analysis: This is a review of a movie. Keywords in this review are twist, classic and not a conventional remake etc. The sentiment of this review is positive. We could see that all three people captured the positive opinion to the movie and keywords in the review. However, the understanding of the review in the first possible summary is a little bit biased. The original review text means that the person Henry winkler is good rather than the story. In addition, there is the grammar error in the second possible summary. We can think that the summary from machine could also contain these errors. The third possible summary is the expected summary from machine.

4.2.2 Qualitative Example 2

Review: This adaptor is real easy to setup and use right out of the box. I had not problem with it at all, it is well worth the purchase. I recommend this adaptor very much for viewing your Nook videos on your HDTV. I just disagree with other reviews on the length of the adaptor, I found it to be fairly adequate as to how and where it is connected to my TV. For me it was just right not too long or too short, I was able to place my Nook right below the connection on the TV stand, it did not fall or anything else, it is fine. Use your own judgement, I'm too busy watching my movies :)

Reference Summary: A Perfdect Nook HD+ hook up

Possible Summary 1: This adaptor with suitable length is easy to set up. I strongly recommend it to view Nook videos on HDTV.

Possible Summary 2: The adapter is convenient and worthwhile. It is suitable for Nook videos on the HDTV. Its length is fine.

Possible Summary 3: This adaptor is easy to setup and has suitable length.

Analysis: This is a review of adaptor which is from the topic electronics. Keyphrases in this review are adaptor, easy to setup, worth, length etc. The sentiment of this review is positive to the product. We can see that all three people captured the positive opinion to the adaptor and keywords in the review. The meanings of the three possible summaries match the meaning of the original review. However, a machine may give us some biased summary. The reason is that some negative words, such as disagree, appear in the original review. In addition, the customer wrote the content which is not related to the adaptor in the end of the review. Furthermore, the original review contains other things, such as Nook, HDTV. Thus, the machine may be confused with the sentiment. On the other hand, the machine could regard Nook or HDTV as the reviewed product. Actually, we found that there exists some noises in our dataset. We can see that the meaning of the reference summary is not consistent with the content of the original review text. We will be careful with the fitting of noises when implementing our research. In other words, we will be careful with the problem of overfitting.

4.2.3 Qualitative Example 3

Review: This is the BEST THANKSGIVING special around...and I am not just saying that because I am the RANKIN/BASS Historian. The sound on this VHS is magnificent and it should be on DVD! I have the original GAS LP Soundtrack on CD. MAURY LAWS and JULES BASS did a wonderful job with the score. ELBOW ROOM, sung by TENN. ERNIE FORD was part of the RANKIN/BASS off-Broadway play A MONTH OF SUNDAYS, before it was put to GREAT use in this special! A GREAT Special! E-mail me with any questions: Rickgoldsc@aol.com Reference Summary: BEST THANKSGIVING special out there!

Possible Summary 1: A RANKIN/BASS Historian recommend the BEST THANKSGIVING VHS very much.

Possible Summary 2: The best Thanksgiving special. The sound is magnificent. I have the original GAS LP soundtrack with good songs on CD, email me.

Possible Summary 3: best vhs as a thanksgiving special.

Analysis: This is a review of a VHS which is from the topic CDs and Vinyl. Keyphrases in this review are BEST THANKSGIVING special, VHS, sound, magnificent etc. The sentiment of this review is positive. We can see that all three people captured the right sentiment and keywords of this review. However, the original review text is mess with so many names and other products, such

as CD, DVD. The machine could be confused with the keywords in the original text.

Reference

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development, vol. 2, no. 2.*

Turney. 1999. Learning to extract keyphrases from text. *Technical report, National Research Council, Institute for Information Technology.*

Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. *In Proceedings of the 16th International Joint Conference on Artificial Intelligence.*

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. *In Proceedings of the 2004 conference on empirical methods in natural language processing.*

Rush, A.M., Chopra, S. Weston, J. 2015. A Neural Attention Model for Sentence Summarization. *The 2015 Conference on Empirical Methods on Natural Language Processing (EMNLP).*

Lin, C. Y. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out.*

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 6572.

Das, S. and Chen, M., 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards.

Hearst, M, 1992. Direction-based Text Interpretation as an Information Access Refinement.

Huettner, A. and Subasic, P., 2000. Fuzzy Typing for Document Management.

Turney, P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.

Pang, B., Lee, L., and Vaithyanathan, S., 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques.

McAuley, J. J., Leskovec, J. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. *In Proceedings of the 22nd international conference on World Wide Web (pp. 897-908). ACM.*

Luong, M. T., Pham, H., Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025.*

Yin, W., Schtze, H., Xiang, B., Zhou, B. 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics, 4, 259-272.*

Zhou, Q., Yang, N., Wei, F., Zhou, M. 2017. Selective encoding for abstractive sentence summarization. *arXiv preprint arXiv:1704.07073.*

Lin, J., Sun, X., Ma, S., Su, Q. 2018. Global encoding for abstractive summarization. *arXiv preprint arXiv:1805.03989.*

Pennington, J., Socher, R., Manning, C. 2014. Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).*

J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. RecSys, 2013.

Chen, Y. C., Bansal, M. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080.*