Sentiment Analysis and Summarization of Restaurant Reviews

Bhuvana Surapaneni
UMass Amherst
bsurapaneni@umass.edu

Janani Krishna
UMass Amherst
jkrishna@umass.edu

Abstract

With emerging increase in the number of food outlets, we can find hundreds and thousand of reviews for each restaurant on the internet. Reading all of these reviews is impossible. Also, one can miss out some relevant reviews about the restaurant and might misjudge the restaurant with fewer reviews. In our project we tried to address this problem by summarizing all the reviews of a restaurant into a single paragraph using bidirectional sequence to sequence models along with attention models.

1 Introduction

In the era of social networking sites, there are a large number of crowdsourced reviews on the internet. The information is also growing in a rapid pace. Especially in the domain of food and dining, many popular websites like Yelp are hosting dedicated web pages with thousands of reviews for every restaurant. It is a very tedious job to go over every review to get the complete information about a business. Hence, consumers usually tend to look at the top most 5 or 10 reviews and make decisions based on those. But, the top reviews may not always give accurate and complete details. This might mislead the consumer to make wrong judgments. Instead, having a concise one paragraph summary of all the thousand reviews will be much helpful. In our project, we tried to create such non-redundant and informative summaries of the Yelp restaurant reviews.

We can easily draw many applications of text summarization in other fields. For instance, search engines can use text summarization techniques to return short summary snippets, which briefs the content of the web search. Also, summarizing news events can save a lot of time for the reader, as the summaries are usually very short compared to original document, yet delivers all the information. In this paper we have outlined our experiments with different variants of recurrent neural networks(RNNs) like sequence to sequence model, bidirectional long short term memory networks(Bidirectional LSTMs) and attention models to create abstractive summaries for the restaurant reviews.

This paper is organized as follows: Section 2 covers all the previous work performed in this area. We have described our dataset in section 3. Our approach to create summaries is presented in section 4. Section 5 gives details about our evaluation strategy. While our results and future work is described section 6 and 7 respectively.

2 Related Work

Over the years there has been a significant work on summarizing massive textual data. Liu et al. (2012) developed a movie rating and review summarization system that focuses on generating feature based summarization using latent semantic analysis(LSA). Primarily, the summaries obtained are essentially dependant on the product feature and opinion mining.

For large documents with repeated information(redundancies), Mittal et al. (2014) has proposed a graph based extractive summarization

Dataset	Size	Unique Business/Products	Unique Users	Reviews Count	Summary count
Amazon Fine Food	300.9 MB	74259	256060	568,454	568,454
Yelp Restaurant	2.46 GB	51613	841403	2927731	0

Table 1: Basic statistics of datasets

technique that focuses on reducing size of the text by eliminating sentences with redundant information. In order to achieve this, a jaccard based similarity score is computed between every pair of sentences. The score describes the extent of word co-occurrence and ability of the sentence to cover other sentences.

Erkan and Radev (2004) came up with LexRank, which produces extractive text summaries for multiple, yet similar documents. In this graph based approach, each sentence in a document are represented as vertices and the edges between these vertices describe the similarity relations. They also proposed various measures to assess the centrality of each sentence in the cluster. These scores are used for extracting the important sentences into the summary.

Denil et al. (2014) proposed an extended convolution neural network, which learns convolution filters at the word, the sentence and the document level, thereby hierarchically learning to capture and compose low level lexical features into high level semantic concepts. This approach overcomes the drawback of N-gram based models by preserving the order of information between words in a sentences and between the sentences in the documents. This paper concentrates more on the extractive summaries rather than the abstractive summaries.

Recently, Yousefi-Azar and Hamey (2017) proposed a ensemble based extractive text summarization technique which uses deep auto encoder and decoders. In this approach the sentences are initially ranked using cosine similarity. A set of random noises are added to input sequence and passed to a ensembled model, which picks the highly weighted sentences as extractive summaries. These picked sentences are highly informative among the input text.

Over the years, the work on text summarization was confined to obtain important sentences as summaries to the text. (Banko et al., 2000) has proposed an alternative to extractive summarization in their paper. They have generated less than a sentence long headlines for news articles using statistical models like term selection, term ordering etc. Rather than using semantically similar words from wide range of vocabulary, the words in the headlines are picked up from the document itself.

The research of (Cohn and Lapata, 2008) led to the development of a framework which finds the summaries of the sentence. These generated summaries convey the necessary and important information with grammatical sense. The authors used a tree-to-tree transduction model which takes cares of structural and lexical mismatches. This model incorporates a novel grammar extraction method and uses a language model for generating coherent results. However, this system tend to replace less common words which leads to counter-intuitive substitution. For instance, if the corpus has fewer samples of the word 'her' when compared to 'his', then all the occurrences of 'her' are replaced by the latter word.

The recent advancements in using recurrent neural networks(RNN) in natural language processing tasks have helped in generating abstractive summaries. Sutskever et al. (2014) have proposed a model for English to French translation. It is known as sequence to sequence model, that is built on a deep LSTM encoder-decoder frame-

work. In the model, one LSTM is used to encode a fixed size input vector, which can be any text like an extract of original language(in the context of language translation), while the other LSTM uses the vector as initial state. Vinyals and Le (2015) applied the same network architecture to create a chat-bot which mimic-ed human conversations. This same idea can also be applied to paraphrase a block of review text. Hu et al. (2015) experimented with a similar RNN network which achieved promising results in the task of summarizing Chinese text.

The above mentioned, standalone LSTM encoder-decoder network may fail in the case where it is given a very long or information-rich text. This is because of the absence of selective encoding and thus the model is forced to encode irrelevant information. But, in the case of text summarization, a high degree of relation exists between the generated text and a particular part of the input text. We can account to the above issue by using attention mechanism. This mechanism allows the decoder to look back to the original input sequence. A context vector is created using the input hidden state sequence, which is used as a condition on the decoder during generating summaries for the text. This approach was first used by Bahdanau et al. (2014) for improving the performance of long sequences in machine translation.

Rush et al. (2015) used a similiar approach for the task of sentence level text summarization. Here, each output word in the summary was conditioned on the original input sequence through an attention mechanism. But the grammar of the generated summaries are inferior. There is a lot of on going research in this area.

3 Data

We have used "Yelp Dataset Challenge reviews dataset"(YelpInc, 2016) for our project. This dataset contains 1,569,264 restaurant reviews.

Compound score	Star points
$>= -1$ and < -0.85	0.5
$>= -0.85$ and < -0.6	1.0
$>= -0.6$ and < -0.45	1.5
$>= -0.45$ and < -0.2	2.0
$>= -2$ and < 0	2.5
$>= 0$ and < 0.2	3.0
$>= 0.2$ and < 0.45	3.5
$>= 0.45$ and < 0.6	4.0
$>= 0.6$ and < 0.85	4.5
$>= 0.85$	5

Table 2: NLTK Vader score to Star mapping

Each review has user information, review text along with star rating for a particular restaurant. But, the dataset has no summary information for training the summarization model. Annotating the dataset in this case would be summarizing the reviews. This task of manual summarizing of all reviews is in itself a hard problem. Therefore, in order to obtain the summary on the procured Yelp dataset we used Amazon fine food dataset(KaggleInc, 2015). This dataset had food product reviews from Amazon. Also, each review has corresponding summary. Since both Yelp and Amazon share a similar context we have trained our model on Amazon fine food dataset, and used this pre-trained model to predict the summaries on Yelp dataset. Some basic statistics of both the datasets are presented in table [1].

4 Method

4.1 Word Embeddings

Word Embeddings (distributional vectors) follow a hypothesis wherein words with similar meanings tend to occur in similar context. These vectors captures the characteristics of the neighboring words. The main purpose of these word embeddings are that they capture similarity between words. Word embeddings are often used

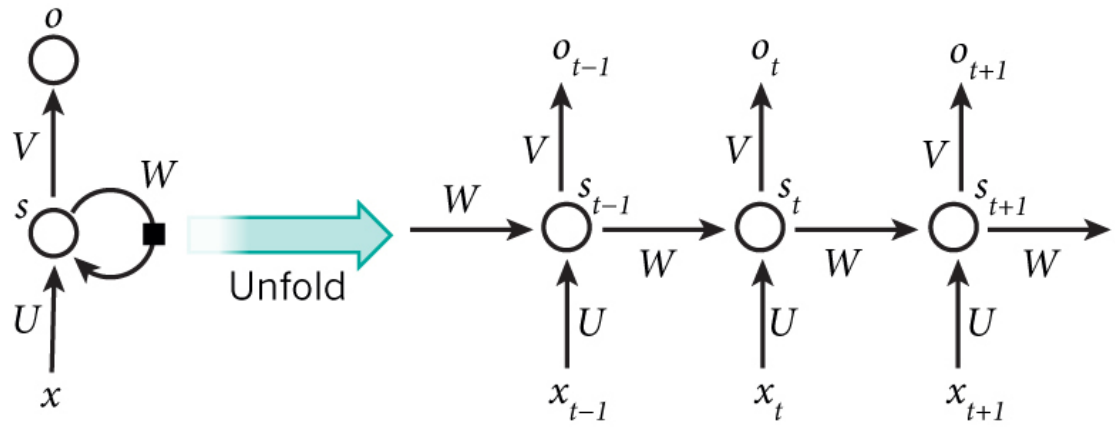

Figure 1: Simple RNN network

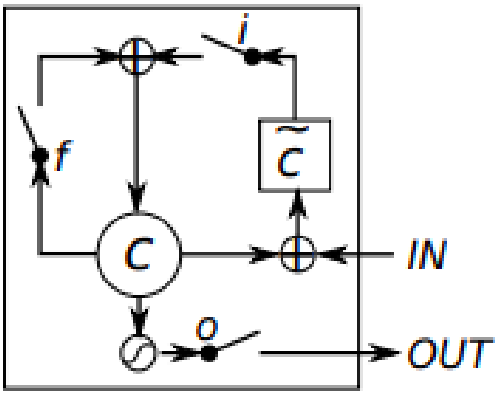

Figure 2: Long short-term memory Model

as the first data processing layer in a deep learning model. Speer and Chin (2016) proposed an approach to represent words as embeddings in a machine-learned vector space. This vector space is generated using an ensemble method that combines embeddings from GloVe (Pennington et al. (2014)), word2vec (Mikolov et al. (2013)) along with ConceptNet(Speer et al. (2016)) and PPDB. Therefore we shifted from Glove to numberbatch. ConceptNet is a knowledge graph that connects words and phrases of natural language with labeled edges.

4.2 Simple Recurrent Neural Networks

Elman network, a three-layer network forms the base for the RNN. In figure[1], x_t is the wordembeddings of the input to the network at time t,o_t is the output of the network,s_t represents the hidden state at time t which is represented as:

$$s_t = f(U_{x_t} + W_{s_{t-1}}) \tag{1}$$

In above equation the s_t is calculated based on the current input and previous hidden state. A nonlinear transformation function f like tanh,ReLU is used and U,V,W are the weights. The hidden state is considered to be the memory unit of the network, that accumulates information from each time steps. Oftentimes, it becomes hard to learn and tune the parameters due to the all-famous vanishing gradient issue. This limitation is overcome by the LSTM models.

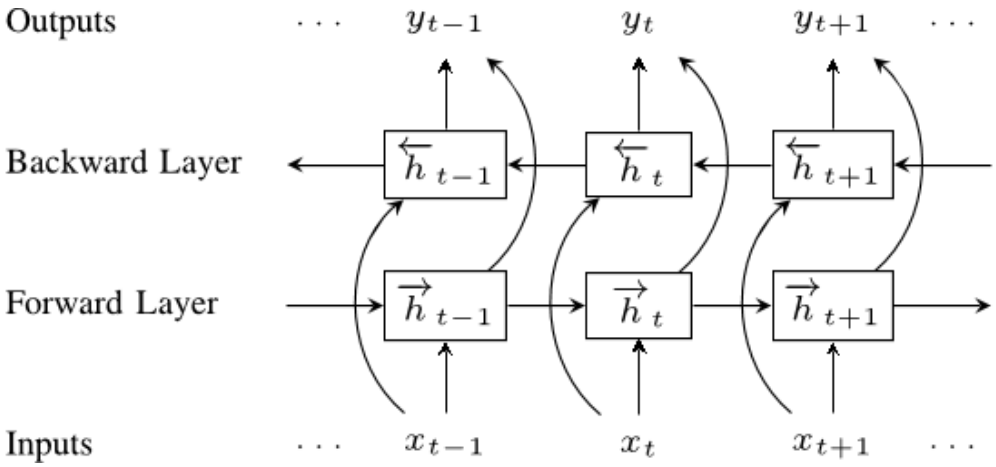

Figure 3: Bidirectional LSTM

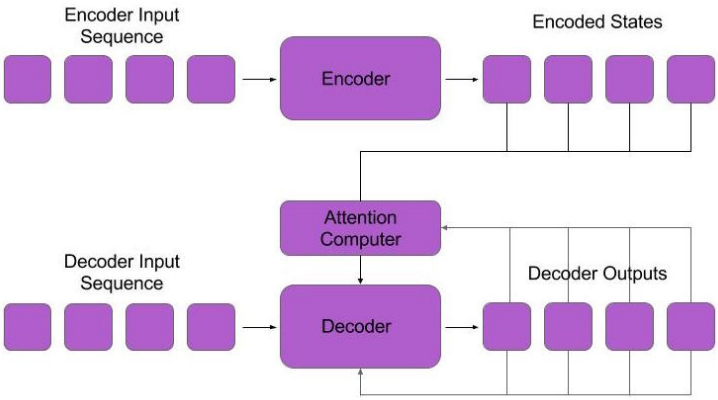

Figure 4: Attention Model

4.3 Long Short Term Memory networks

Hochreiter and Schmidhuber (1997),came up with a RNN like architecture in conjunction with a gradient-based learning algorithm. This network has additional gates called "forget" gates over the simple RNN. An LSTM block depicted in figure[2], primarily contains four components: a cell, an input gate, an output gate and a forget gate. The cell plays the role of memory, remembering values over every time step. Each of these gates can be considered as a neuron in a multi-layer neural network.Each j-th LSTM unit maintains a memory c_t^j at time t.The output h_t^j of the LSTM is given as:

$$h_t^j = o_t^j tanh(c_t^j) \tag{2}$$

here o_t^j is an output gate that controls the amount of memory content exposures. The output gate is defined as:

$$o_t^j = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t)^j \tag{3}$$

where σ is sigmoid function. V_o is a digonal matrix. The memory cell c_t^j is updated by partially forgetting the existing memory and updating a new memory content :

$$\bar{c}_t^j = tanh(W_c x_t + U_c h_{t-1})^j \tag{4}$$

The forget gate f_t^j controls the extent with which the existing memory is forgotten while the input gate i_t^j controls the degree to which new memory is added to the memory cell.These gates are com-

puted as below:

$$f_t^j = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1})^j \quad (5)$$

$$i_t^j = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1})^j \quad (6)$$

where V_f and V_j are diagonal matrices. In order to have the short-term information for longer time, we can memorize the information from both the ends of the input sequence. This is achieved using Bi-directional LSTMs.

4.4 Bi-directional LSTMs

Schuster and Paliwal (1997) introduced bidirectional LSTM/RNN to increase the amount of input information available to the network.The basic concept of bidirectional model is to connect two hidden layers of opposite directions to the same output.The output layer can thus get information from past and future states. In principle the bidirectional model as depicted in figure[3] splits the neurons of a regular RNN into two directions, one for forward state and other backward state. We have used this model to generate the sequence of words for the given review text.

4.5 Attention Mechanism

Attention is the idea of freeing the encoder-decoder architecture from the fixed-length internal representation as proposed by Bahdanau et al. (2014).This mechanism is achieved by keeping the intermediate outputs from the encoder LSTM from each step of the input sequence and training the model to learn to pay selective attention to these inputs and relate them to items in the output sequence. The figure[4] depicts the attention model where the decoder receives the attention model's state, that is the weighted average of the encoder states.

5 Evaluation

We need to have concrete evaluation strategy to measure the accuracy of the summaries generated

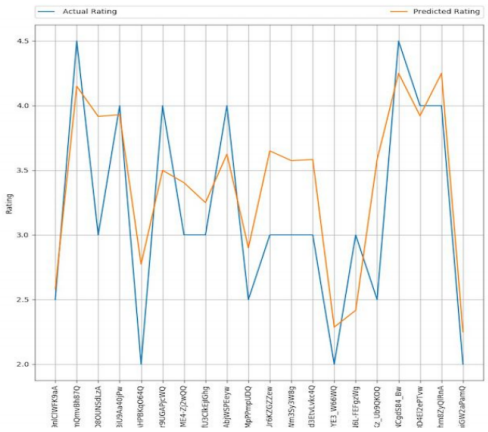

Figure 5: Analyzing VADER performance on Yelp data

by both the techniques. One measure of evaluation can be human judgment. We can go over few reviews and the corresponding generated summaries and evaluate the validity. But this is often subjective and time consuming task. Therefore, we have used the following two techniques to evaluate our methods.

5.1 Entropy Loss

The definition of cross entropy in information theory between two probability distributions p and q measures the average number of bits needed to identify an event drawn from the set. In short entropy measures the average information content. We are more curious about how this measure, for the generated images, changes through multiple epochs. In our project we have used the weighted average of the entropy loss which we average over a set of batches.

5.2 Sentiment Analysis using Vader

NLTK Sentiment Analyzer uses VADER algorithm proposed by Hutto and Gilbert (2014), which analyzes a piece of text and gives scores to the following four classes: positive, negative, neutral and compound. The compound score ranges from [-1, 1]. We have divided this score into different range bins, and assigned a star rating to each

Original Review	Tried many varieties of BBQ chips and hands down these are the BEST on the market! These R not too salty or greasy. Great crispy crunch. Plenty of good BBQ flavor with a hint of sweetness that all other brands fall short of. Every bag (after bag) is just as mouth wateringly delicious as the 1st!!'
Original Summary	BEST BUY in BBQ Chips
Without Attention	best barbeque buy in barbeque barbeque
With Attention	best taste BBQ buy

Figure 6: Generated Summaries on Amazon dataset

Original Review	Super clean restaurant and friendly staff. FRESH food. Hasn't been sitting under heat lamps. NO MSG, this is the good stuff. I have to have the Kung Pao Chicken weekly'
Without Attention	refreshingly refreshingly refreshingly refreshingly
With Attention	great refreshingly refreshingly

Figure 7: Generated Summaries on Yelp dataset

range as specified in table[2]. As a part of our evaluation technique, we are calculating score on both original review and generated summaries. If the difference in star rating is beyond 1.5, then it is considered as a false summary. We are obtaining average success rate and producing that as our sentiment analysis accuracy.

In order to justify the usage of this evaluation technique, we have used the same technique on Yelp reviews and predicted star rating for all the reviews in the dataset. Then, for each review we have compared this predicted rating with the gold label star rating given by the customer. That gave us 87.2008989983% accuracy. The line plot depicting the actual vs predicted rating for randomly selected business IDs from Yelp data is presented in figure[5]. This results justifies our strategy of using VADER for evaluation on this dataset.

Dataset	Method	Accuracy
Amazon	Without Attention	66.48%
Amazon	With Attention	71.47%
Yelp	Without Attention	32.45%
Yelp	With Attention	33.38%

Table 3: Evaluation results using NLTK Vader

6 Results

An extract of summaries produced by bi-directional LSTM models with/without attention are presented in figures[6] and [7]. As we can observe, the summaries obtained using bi-directional attention models outperform the summaries generated by the model without attention. Also, the summaries generated on Amazon fine food dataset are quite good when compared to that of Yelp summaries. One reason for this behaviour might be the domain of the datasets. Even though, both the datasets have the reviews regarding the food, Amazon dataset contains reviews for the food products sold on Amazon E-commerce site. Whereas, Yelp restaurant reviews are about the dining experience. Yelp reviews may discuss about a wider spectrum of topics like the restaurant ambience, parking availability, seat reservations, employee behaviour etc. This data is missing from our training dataset.

The cross entropy loss for both the models on training is depicted in figures [8] and [9]. As we can observe the average loss for the bidirection

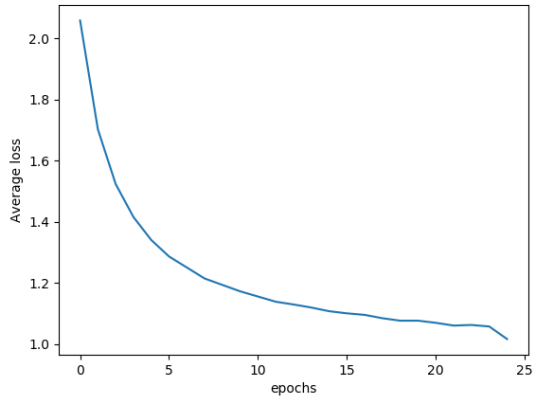

Figure 8: Bidirectional LSTM: Epoch vs Average Entropy Loss

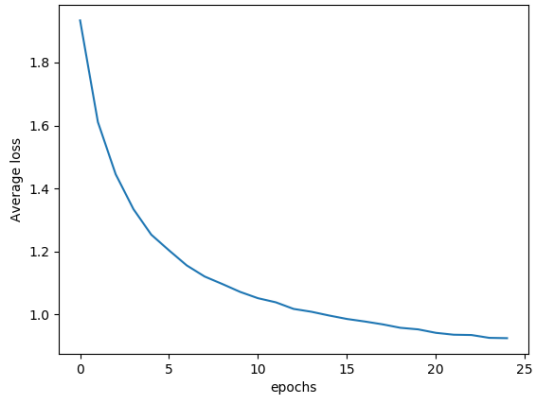

Figure 9: Bidirectional LSTM with Attention: Epoch vs Average Entropy Loss

LSTM with attention over every epoch is reaching the local optima sooner.

For further evaluation, we have used NLTK VADER sentiment analyzer. The results obtained on train and test set are described in table[3]. As we can see, adding attention mechanism to the bi-directional models improved the performance of the text summarization.

7 Discussion and Future Work

Text summarization for reviews is helpful for consumers to get a complete information about restaurants in a very less time. Considerable about of research has been done in the area of text summarization. But, a lot of research is dedicated to extractive summarization. Getting high quality results using abstractive text summarization is still an unconquered research area.

In this project, we have attempted to create summaries on Yelp dataset(which does not have gold label summaries) using Amazon fine food dataset. Even after using state of the art techniques, we have not got satisfactory results. Using a more contextually similar dataset for training the model would have had given better results. But, in general it is hard to find datasets which share the exact same context. In future, we will try to improve the results by fine tuning the model, which is created using Amazon dataset. This can be done by annotating few thousands of reviews on Yelp manually and using this data to tune the hyper parameters of the pre-trained model.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473. http://arxiv.org/abs/1409.0473.

Michele Banko, Vibhu O. Mittal, and Michael J. Witbrock. 2000. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '00, pages 318–325. https://doi.org/10.3115/1075218.1075259.

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '08, pages 137–144. http://dl.acm.org/citation.cfm?id=1599081.1599099.

Misha Denil, Alban Demiraj, Nal Kalchbrenner, Phil Blunsom, and Nando de Freitas. 2014. Modelling, visualising and summarising documents with a single convolutional neural network. *CoRR* abs/1406.3830. http://arxiv.org/abs/1406.3830.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.* 22(1):457–479. http://dl.acm.org/citation.cfm?id=1622487.1622501.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale chinese short text summarization dataset. *CoRR* abs/1506.05865. http://arxiv.org/abs/1506.05865.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. https://www.aaai.org/ocs/index.php/ICWSM/.

KaggleInc. 2015. Amazon fine food reviews.

C. L. Liu, W. H. Hsaio, C. H. Lee, G. C. Lu, and E. Jou. 2012. Movie rating and review summarization in mobile environment. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42(3):397–407. https://doi.org/10.1109/TSMCC.2011.2136334.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781.

Namita Mittal, Basant Agarwal, Himanshu Mantri, Rahul Kumar Goyal, and Manoj Kumar Jain. 2014. Extractive text summarization .

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *CoRR* abs/1509.00685. http://arxiv.org/abs/1509.00685.

M. Schuster and K. K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681. https://doi.org/10.1109/78.650093.

Robert Speer and Joshua Chin. 2016. An ensemble method to produce high-quality word embeddings. *CoRR* abs/1604.01692. http://arxiv.org/abs/1604.01692.

Robert Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR* abs/1612.03975. http://arxiv.org/abs/1612.03975.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pages 3104–3112. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR* abs/1506.05869. http://arxiv.org/abs/1506.05869.

YelpInc. 2016. Yelp open dataset. `https://www.yelp.com/dataset/`.

Mahmood Yousefi-Azar and Len Hamey. 2017. Text summarization using unsupervised deep learning. *Expert Systems with Applications* 68(Supplement C):93 – 105. https://doi.org/https://doi.org/10.1016/j.eswa.2016.10.017.