W266 Final Project
Fall 2018
David He, Rory Liu & Andrea Sy

Utilizing Abstractive Text Summarization on Amazon Food Reviews & WikiHow Articles

Overview

In this day and age, there is a culture of “data openness” that is starting to take hold, and an abundance of data is rapidly being made available to the general public. This type of data can take on many forms, from collections of product and movie reviews to vast libraries of historical documents. Given the volume of text that is now available, the rise in popularity of text summarization comes as no surprise.

The team decided to focus on automatic and abstractive text summarization because of its significance and the many ways a model focusing on this could be applied. In this paper, the team will go through the dataset selected, the general goal and approach to creating the text summaries, as well as the results obtained from the models built.

Abstract

As mentioned in our overview, the ubiquity of data has rendered manual text summarization inefficient and unrealistic. Upon assessing the state of automatic summarization further, the team came to a decision to focus on abstractive summarization as opposed to the extractive, choosing to focus on what can be likened to paraphrasing.

There has been a number of advancements in abstractive summarization, like Nallapati, Zhou, Gulcehre, Xiang (2016) enhancing seq2seq models for text summarization, and See, Liu, Manning (2017) combining extractive techniques and abstractive seq2seq techniques. Although these papers have their own techniques to build on previous iterations of text summarization methodologies, we decided to build our implement our own model with similar concepts. We were also curious about how the model would work on interesting dataset besides news corpus.

We built a sequence to sequence model with attention, and used an industry standard metric called Recall-Oriented Understudy for Gisting Evaluation (ROUGE) to assess the overall accuracy of the model. It works by comparing an automatically produced summary or translation against a set of reference summaries and measuring their overlaps.

We trained and tested this seq2seq model on 2 separate datasets: the Amazon Fine Food Review data, and the WikiHow articles dataset. In doing so, we were keen on understanding how such a model would perform on both creating very short summaries (4-5 words) and slightly longer summaries (6-9 words). We used the same model in the two datasets, but did training separately to make sure we have an apples to apples comparison. The methodology section will focus on both the model used as well as the significance of the ROUGE metric. It is important to note that these datasets are very different in nature to previous datasets assessed in other research papers. While majority of previous research projects focused on objective news articles, the team’s focus on reviews and wikihow articles were very different in tone and objectivity.

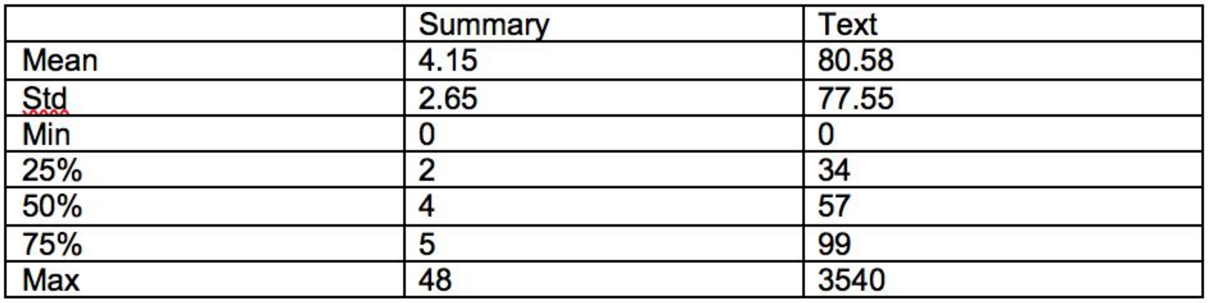
Datasets

Amazon Fine Foods Review Dataset

To begin our training, we first selected the Amazon Fine Food Review data. This dataset contains consumer reviews for fine-food purchases on Amazon, with the full review serving as text body, and the title serving as summary. The data set has nearly 400,000 records, and we used a subset in our training.

Below is the length distribution of the summaries and the texts in this dataset: On average, we have 4-word summaries, and 80-word texts. During data cleaning, we removed records with excessively long or short texts / summaries to reduce padding.

Length Distribution

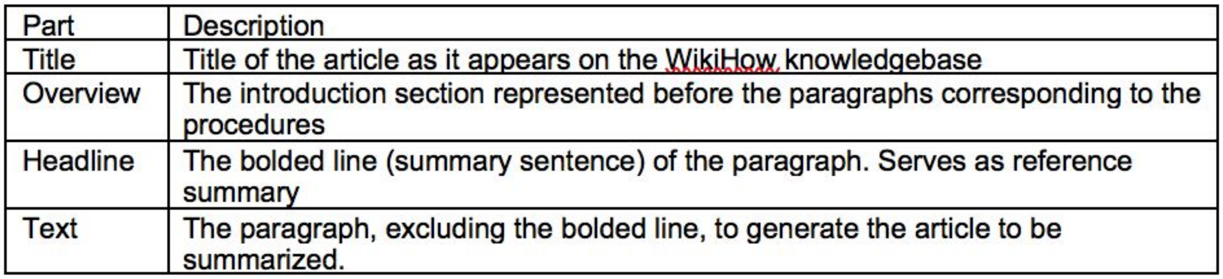


	Summary	Text
Mean	4.15	80.58
Std	2.65	77.55
Min	0	0
25%	2	34
50%	4	57
75%	5	99
Max	48	3540

WikiHow Dataset

For our second dataset, we chose a recent WikiHow articles compilation (updated October 2018). According to the resource that we pulled the dataset from, each article is comprised of several paragraphs. Each paragraph has a main sentence that acts as a summary. Through merging, the result is a dataset that has 1.37million rows, each consisting of 1 main summary sentence, as well as the paragraph without this main sentence.

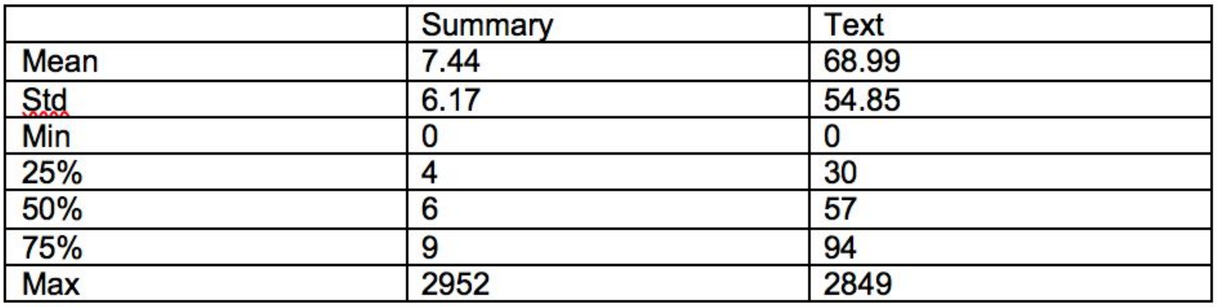
Below you will find a table that presents the components and a brief description of its contents.



Part	Description
Title	Title of the article as it appears on the WikiHow knowledgebase
Overview	The introduction section represented before the paragraphs corresponding to the procedures
Headline	The bolded line (summary sentence) of the paragraph. Serves as reference summary
Text	The paragraph, excluding the bolded line, to generate the article to be summarized.

Below is the length distribution of both the summary and text of this dataset: as mentioned, this dataset contains slightly longer summaries, but similar lengthed text compared to our first amazon reviews dataset. Similar to our approach with the Amazon set, we removed excessively long or short texts / summaries in our training.

Length Distribution



	Summary	Text
Mean	7.44	68.99
Std	6.17	54.85
Min	0	0
25%	4	30
50%	6	57
75%	9	94
Max	2952	2849

Methodology

For the purposes of this project, the team underwent the standard process of assessing and cleaning of each dataset, defining the parameters for training the data, building out and running a model before evaluation with an industry standard metric. The team trained the model on both datasets before testing them respectively.

I. Assessment and Cleaning

Standard processes were applied to both the Amazon Fine Food and WikiHow datasets, including dropping "NA" or empty cells and duplicates that could be found. Once that had been completed, the team then sought to cover all bases by adding a list of contractions and a corresponding expanded definition. Additional cleaning was done to remove unwanted symbols, unnecessary stops to prevent fewer nulls in embeddings. The final step was to get an idea of the general length of these WikiHow texts.

II. GloVe and Text Conversion

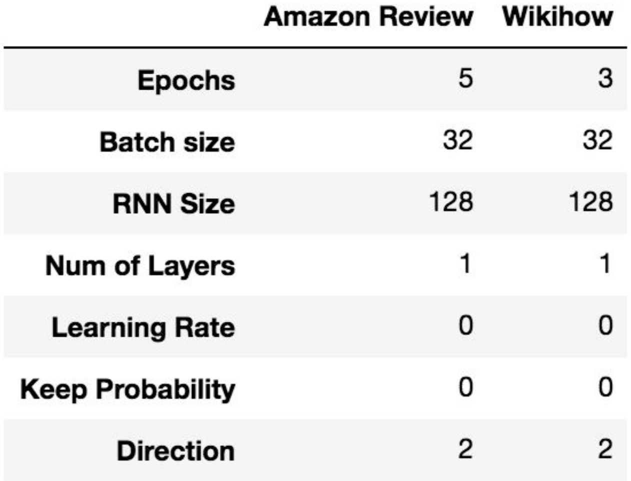
The team utilized GloVe in order to convert words into vector forms since the model that the team utilized will interpret the generated numbers instead of the text. The model built in the functionality to select from a list of available embedding dimensions (50, 100, 200, 300). For our final training run we used the 200d embedding because we found that this was the best balance between training time and performance. Instead of converting all words to embeddings up front, the team decided to do the embedding in the model's embedding layer to improve efficiency.

III. Parameters & Training

To generate abstractive summarization model, the team opted to use a bi-directional RNN for the encoding portion, and attention mechanism for the decoding layer. Other logistical factors that the team took into account was sorting these reviews by length. Sorting them from shortest to longest ensured that similar length texts would be batched together, reducing the need for padding.

To build out the training and inference decoding layers as mentioned above, Tensorflow functions focused on reading sequences of integers, processing and output creation were utilized. Additionally, the team set certain hyperparameters for the training of this model. After iterating on the hyperparameters for the sake of efficiency, these were the final criteria agreed upon.

Hyperparameters for the Datasets



	Amazon Review	Wikihow
Epochs	5	3
Batch size	32	32
RNN Size	128	128
Num of Layers	1	1
Learning Rate	0	0
Keep Probability	0	0
Direction	2	2

It is important to note that computer power is a limitation that factored into the decided hyperparameters. In an ideal world, the more epochs trained the better. Due to the constraints, it was unrealistic to test various hyperparameters with multiple epochs in each round.

Once the hyperparameters were set, the team proceeded to train the model on a subset of data. In the following portion, the team added a section where summaries could be generated from your own descriptions or something that you could derive from the data. One key point here is that we maintained the range of summary lengths for consistency. Once the model was complete, and we obtained results from the model, the team proceeded to evaluate using the ROUGE metric.

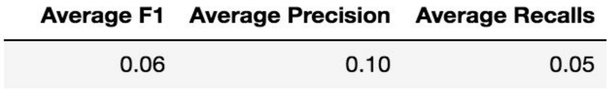
Results & Evaluation

For our evaluation purposes, the team opted to utilize an industry standard, the ROUGE metric. The code implementation that we opted for yields a rouge-n metric, which helps us evaluate how many groups of n-grams were present in both a human summary and the summary generated by the model.

The team obtained the rouge-1, rouge-2 and rouge-l scores for one-off generated summaries. Additionally, we obtained a rouge-1 metric for a random sample of 100 generated summaries and obtained the mean for the f1, precision and recall scores to obtain a benchmark for how our model performed. The reason we mostly focused on aggregate ROUGE-1 over ROUGE-2 or other finer granularity ROUGE measures, is that we mostly cared about whether same general concepts are discussed between an automatic summary and a reference summary, especially when the automatic summary is concise.

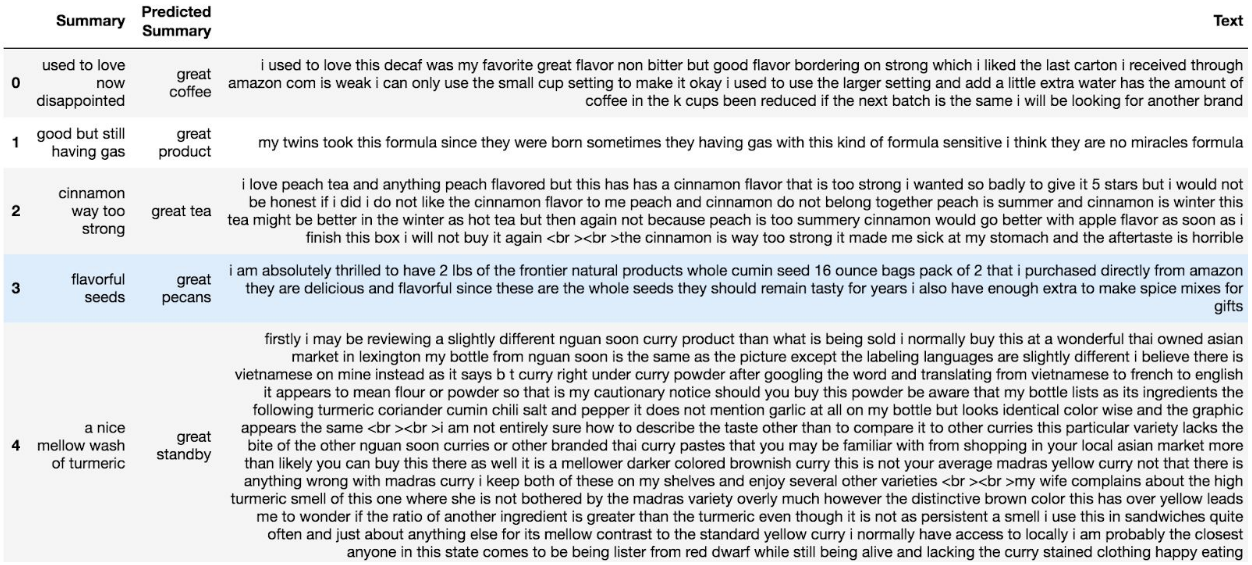
You can find the results below.

Amazon Review ROUGE-1 Scores



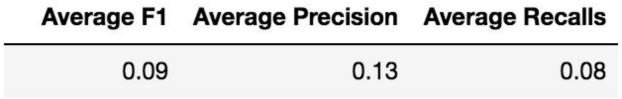
Average F1	Average Precision	Average Recalls
0.06	0.10	0.05

Amazon Generated vs Predicted Summaries



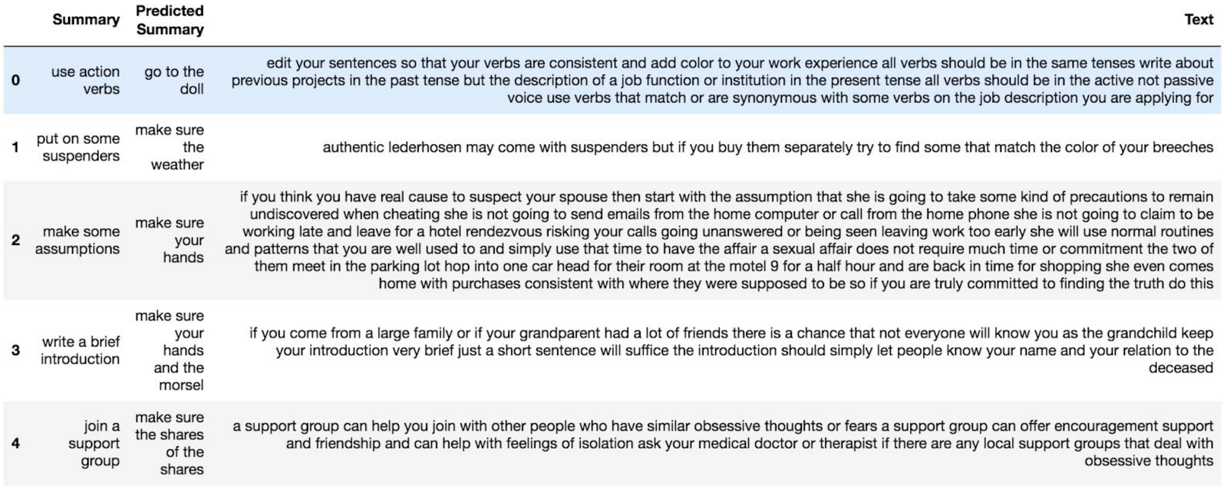
	Summary	Predicted Summary	Text
0	used to love now disappointed	great coffee	i used to love this decaf was my favorite great flavor non bitter but good flavor bordering on strong which i liked the last carton i received through amazon com is weak i can only use the small cup setting to make it okay i used to use the larger setting and add a little extra water has the amount of coffee in the k cups been reduced if the next batch is the same i will be looking for another brand
1	good but still having gas	great product	my twins took this formula since they were born sometimes they having gas with this kind of formula sensitive i think they are no miracles formula
2	cinnamon way too strong	great tea	i love peach tea and anything peach flavored but this has has a cinnamon flavor that is too strong i wanted so badly to give it 5 stars but i would not be honest if i did i do not like the cinnamon flavor to me peach and cinnamon do not belong together peach is summer and cinnamon is winter this tea might be better in the winter as hot tea but then again not because peach is too summery cinnamon would go better with apple flavor as soon as i finish this box i will not buy it again <br ><br >the cinnamon is way too strong it made me sick at my stomach and the aftertaste is horrible
3	flavorful seeds	great pecans	i am absolutely thrilled to have 2 lbs of the frontier natural products whole cumin seed 16 ounce bags pack of 2 that i purchased directly from amazon they are delicious and flavorful since these are the whole seeds they should remain tasty for years i also have enough extra to make spice mixes for gifts
4	a nice mellow wash of turmeric	great standby	firstly i may be reviewing a slightly different nguan soon curry product than what is being sold i normally buy this at a wonderful thai owned asian market in lexington my bottle from nguan soon is the same as the picture except the labeling languages are slightly different i believe there is vietnamese on mine instead as it says b t curry right under curry powder after googling the word and translating from vietnamese to french to english it appears to mean flour or powder so that is my cautionary notice should you buy this powder be aware that my bottle lists as its ingredients the following turmeric coriander cumin chili salt and pepper it does not mention garlic at all on my bottle but looks identical color wise and the graphic appears the same <br ><br >i am not entirely sure how to describe the taste other than to compare it to other curries this particular variety lacks the bite of the other nguan soon curries or other branded thai curry pastes that you may be familiar with from shopping in your local asian market more than likely you can buy this there as well it is a mellower darker colored brownish curry this is not your average madras yellow curry not that there is anything wrong with madras curry i keep both of these on my shelves and enjoy several other varieties <br ><br >my wife complains about the high turmeric smell of this one where she is not bothered by the madras variety overly much however the distinctive brown color this has over yellow leads me to wonder if the ratio of another ingredient is greater than the turmeric even though it is not as persistent a smell i use this in sandwiches quite often and just about anything else for its mellow contrast to the standard yellow curry i normally have access to locally i am probably the closest anyone in this state comes to be being lister from red dwarf while still being alive and lacking the curry stained clothing happy eating

Wikihow Dataset ROUGE-1 Scores



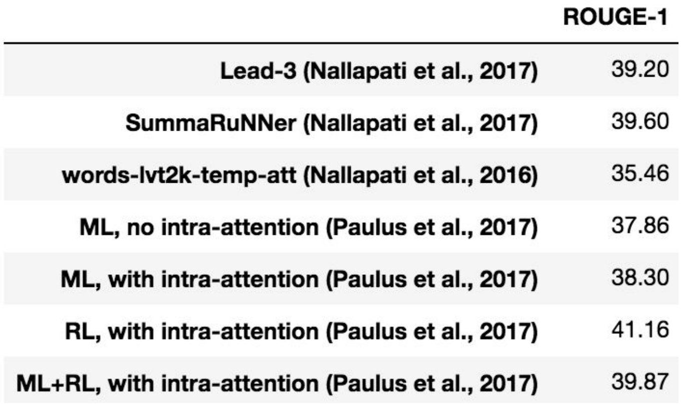
Average F1	Average Precision	Average Recalls
0.09	0.13	0.08

WikiHow Generated vs Predicted Summaries



	Summary	Predicted Summary	Text
0	use action verbs	go to the doll	edit your sentences so that your verbs are consistent and add color to your work experience all verbs should be in the same tenses write about previous projects in the past tense but the description of a job function or institution in the present tense all verbs should be in the active not passive voice use verbs that match or are synonymous with some verbs on the job description you are applying for
1	put on some suspenders	make sure the weather	authentic lederhosen may come with suspenders but if you buy them separately try to find some that match the color of your breeches
2	make some assumptions	make sure your hands	if you think you have real cause to suspect your spouse then start with the assumption that she is going to take some kind of precautions to remain undiscovered when cheating she is not going to send emails from the home computer or call from the home phone she is not going to claim to be working late and leave for a hotel rendezvous risking your calls going unanswered or being seen leaving work too early she will use normal routines and patterns that you are well used to and simply use that time to have the affair a sexual affair does not require much time or commitment the two of them meet in the parking lot hop into one car head for their room at the motel 9 for a half hour and are back in time for shopping she even comes home with purchases consistent with where they were supposed to be so if you are truly committed to finding the truth do this
3	write a brief introduction	make sure your hands and the morsel	if you come from a large family or if your grandparent had a lot of friends there is a chance that not everyone will know you as the grandchild keep your introduction very brief just a short sentence will suffice the introduction should simply let people know your name and your relation to the deceased
4	join a support group	make sure the shares of the shares	a support group can help you join with other people who have similar obsessive thoughts or fears a support group can offer encouragement support and friendship and can help with feelings of isolation ask your medical doctor or therapist if there are any local support groups that deal with obsessive thoughts

Other Research Papers (On CNN/Daily Mail articles dataset)



	ROUGE-1
Lead-3 (Nallapati et al., 2017)	39.20
SummaRuNNer (Nallapati et al., 2017)	39.60
words-lvt2k-temp-att (Nallapati et al., 2016)	35.46
ML, no intra-attention (Paulus et al., 2017)	37.86
ML, with intra-attention (Paulus et al., 2017)	38.30
RL, with intra-attention (Paulus et al., 2017)	41.16
ML+RL, with intra-attention (Paulus et al., 2017)	39.87

Conclusion & Recommendations

The model and ROUGE metrics yielded very interesting results. For instance, the Amazon reviews generated summaries that may not have yielded a word to word match, but some of the generated summaries were still able to capture the gist of the text pulled from the dataset. You can see a sample set of the Amazon reviews above.

Upon further assessment of the results of our model, the training set that the team used for both WikiHow and Amazon Fine Food Reviews datasets were about 10,000 rows in size. Increasing that and leaving the model to train for longer (more epochs) could improve the accuracy of our model and the summaries that could be generated.

In conclusion, based on comparing our ROUGE scores against benchmarks, utilizing a model to train on multiple datasets is not the most effective approach for abstractive text summarization. There are certain nuances that different types of texts have that cannot be captured in a one-size fits all model. Certain parameters and modifications still need to be made to adapt to different types of data. There is still a lot of work that needs to be done in the field of NLP and text summarization before models like these can be applied universally to the texts available.

Bibliography

Romain Paulus, Caiming Xiong, Richard Socher. 2017 A Deep Reinforced Model For Abstractive Summarization

Abigail See, Peter J. Liu, Christopher D. Manning. Get To The Point: Summarization with Pointer-Generator Networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp.1073–1083, 2017

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization ¨ using sequence-to-sequence rnns and beyond. Proceedings of SIGNLL Conference on Computational Natural Language Learning, 2016.

Sumit Chopra, Michael Auli, Alexander M Rush, and SEAS Harvard. Abstractive sentence summarization with attentive recurrent neural networks. Proceedings of NAACL-HLT16, pp. 93–98, 2016.

Vishal Gupta, Gurpreet Singh Lehal. A Survey of Text Summarization Extractive Techniques. Journal of Emerging Technologies in Web Intelligence, Vol 2, No.3, August 2010.