Contextual Linking of News Articles

Aamir Rasheed (aamirar@stanford.edu) and Dunia Hakim (dunia@stanford.edu)
CS 230 Deep Learning, Stanford University

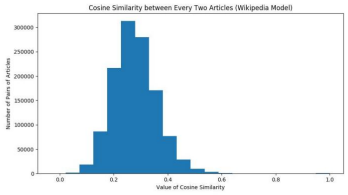
Abstract

To mitigate the issue of misinterpreting events and information due to a lack of contextual information in news stories, our project aims to link new, unseen news articles to related articles in a small dataset of 1224 news stories. To do this, we used a doc2vec model trained on a large wikipedia corpus to generate vector embeddings for our dataset. For our model, we used a clustering approach from [1] where we construct a sparse weighted graph of cosine similarities using the MST-KNN method, then use a Markov stability algorithm to generate the final clusters. While some of our clusters were extremely good, we were limited by the size of our dataset.

Dataset and Features

Our dataset is a news article database that are unique in that they are bare-bones facts of what is being reported, and are therefore generally short. The average word length was 362. Topics covered were usually world/national news, with very little reporting in arts & entertainment..

For our features, we used a doc2vec model trained on a wikipedia dataset of 35M articles to generate word embeddings for each of our dataset items. However, due to the short length of the articles and little variety in topics covered, many of the vectors were equally related to each other, as pictured in the cosine similarity histogram below.

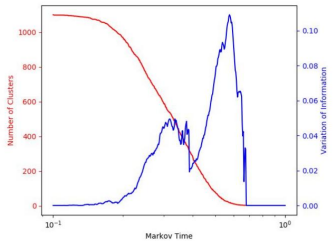


Models and Results

For our baseline approach, we used a pretrained Doc2Vec model that was trained with the algorithm *distributed bag-of-words* on the English Wikipedia corpus. Using this pretrained model and the gensim package, we inferred a vector embedding for the content of each of our articles (each vector embedding had 300 features). We then calculated the cosine similarity between every two vector embeddings and thus were able to get the most contextually related articles for each article in the dataset. While most of our results were quite good, on rare occasions the results seemed unrelated to the target article. Here are some examples of good and bad results:

Example of Good Results	Example of Bad Results
Most similar articles to "US to impose tariffs on $200 billion of Chinese imports":	Most similar articles to "Q&A: A breakdown of the US immigration court system":
#1: "US, China enact new trade tariffs"	#1: "Mueller reportedly subpoenaed Trump financial records from German bank"
#2: "USTR announces 25% tariffs on $50B worth of Chinese goods"	#2: "AT&T's potential Time Warner acquisition delayed"
#3: "Trump, EU leader Juncker agree to 'work towards' zero tariffs, subsidies on non-auto goods"	#3: "AT&T says 'association' with Cohen was 'serious misjudgment"

For our main model, we used the clustering approach from [1]. First, we generated a weighted cosine similarity graph from our vector embeddings. Then, we generated the MST-KNN graph to develop a more sparse representation. Finally, we use Markov Stability to extract multi-scale community structure. Result below.



Dips in variation of information indicate robust clusters. Unfortunately, we only had one dip, at ~300 clusters, and only 7 of those clusters had more than 7 stories. Of those, only one was poorly clustered, represented below in the first word cloud.

Example of bad cluster. None of the key words in this story are related. Of the seven stories in this cluster, only two were related.

Example of a good cluster. We can see this cluster contains stories pertaining to Trump's trade war with China. 6 of the 7 clusters were like this.

Discussions and Future Work

For our baseline approach, the results were quite good. Sadly, the method is extremely slow when dealing with big datasets. As for our clustering approach, we did not achieve our aim of having hierarchical clustering, because only we had only one robust clustering. Even in that clustering, only 7 clusters had more than 7 stories, so it did not help our overall goal of achieving contextual linking to news articles. We speculate that this was mainly because our news article dataset was so small, and that most articles with the exception of a few were on different topics anyways.

If we were to continue working on this project, we would first try using a model that is explicitly trained on news articles. Wikipedia tends to have a large variety of topics discussed whereas the articles in our dataset mostly discuss world news and politics. This may have caused the cosine similarities to be inaccurate relative to the dataset (most similarities were within the range 0.1 and 0.5).

Secondly, we could make use of other features of our dataset to generate better clusters such as the date of the article and the region that it talks about.

Finally, we would augment the dataset and make it large enough in order for the clustering method to work better. Once the clustering is good enough, we would use a two or three hidden layer neural network to classify which cluster an article would be best suited for. This would allow for a faster and more accurate contextual linking than our baseline method.

References

[1] Altuncu, M. Tarik, et al. "Content-Driven, Unsupervised Clustering of News Articles through Multiscale Graph Partitioning." ArXiv:1808.01175 [Cs, Math], Aug. 2018. arXiv.org, http://arxiv.org/abs/1808.01175.