News Clustering and Summarization

Saumya Didwania, Jonathan Lu, Saumyaa Shah
sd4469, jxl219, sns9906

Center for Data Science
New York University

Abstract

- There is a flood of information and a lot of redundancy in news articles published every day. The objective of this project is to comb through all that information and create a summary for each newsworthy event.
- These summaries could provide an abstract of daily events, create a hub where someone can access the summary of different sources' coverage of that event, or even translate an event into another language to allow for global updates.

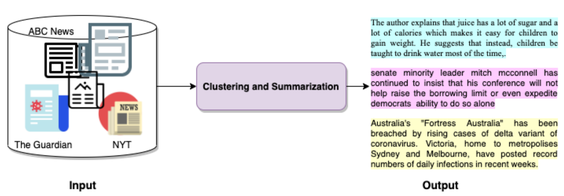


Figure: Process Overview

- To generate a summary from multiple news sources, we propose a method that clusters the articles based on topic and generates a concise summary for each cluster.

Introduction

- **Data:** To obtain the data used to train the model we scraped newspapers from a variety of sources, such as the **Guardian, the New York Times and Wall Street Journal** along with getting articles from a few different smaller sources on **NewsAPI**. Most of our data was from Fall 2021, though some of our news sources had data for the entire year.
- **Algorithms Evaluated:**
  - Embedding: For representing the articles in vector space, we generate embeddings on a **word(Word2Vec)** and **sentence(SentenceBERT)** level.
  - Clustering: To cluster the articles based on their vector representation, we employ **k-means**, **DBSCAN**, **spectral clustering**, etc.
  - Summarization: Finally, to generate the summary, we compare between **extractive summarization(TF-IDF)** and **abstractive summarization(BART)**.
- **Related Work:**
  - The majority of current approaches use event-based clustering around specific topics using both news articles and social media platforms. For example, in Carta et al. [1], the authors use event-based detection for future stock price increases.
  - For summarization, typically abstractive text summarization is used, as they tend to convey more semantic meaning when generating words, rather than looking at 'the numbers' in an extractive approach.

Proposed Methodology

The proposed algorithm consists of 4 stages: **Named Entity Recognition, Embedding, Clustering** and **Summarization**.

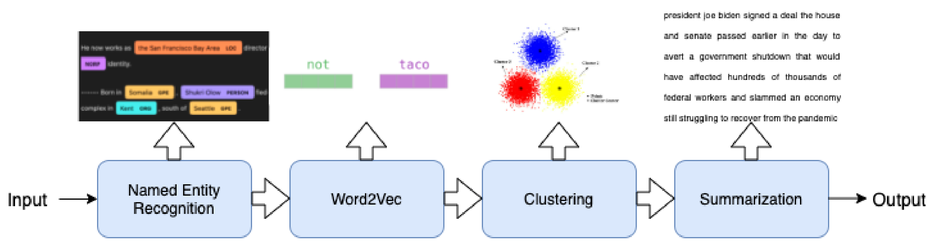
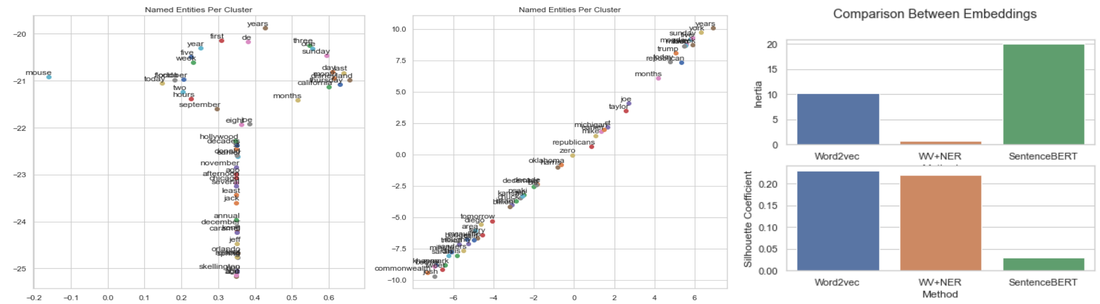


Figure: Detailed Process Overview

- **NER:** Firstly, we extract named entities i.e. entities representing person, place, date/time, etc. from the data. Before generating the embeddings, we filter out articles having less than 2 named entities.
- **Embedding:** Using the named entities, we train a Word2Vec language model to generate embedding vectors for the articles.
- **Clustering:** We apply mini-batch k-means clustering on the embeddings to obtain clusters of similar articles.
- **Summarization:** We feed all the articles in a cluster to BART to generate a concise summary per-cluster.

Results

- **Named Entity Recognition + Clustering:** The first figures represent the **t-sne** visualization of all the unique named entities present in the articles of a cluster.



Observing the inertia and silhouette score for various word-level and sentence-level embeddings, we observe that for named entity embeddings, we get the lowest inertia for a comparable silhouette coefficient.

- **BART Summarization**

Australia's "Fortress Australia" has been breached by rising cases of delta variant of coronavirus. New South Wales and Victoria, home to metropolises Sydney and Melbourne, have posted record numbers of daily infections in recent weeks. The government has struck deals with other countries, including Britain and Singapore, to secure Pfizer doses earlier. Fewer than 35% of Australians are fully vaccinated, putting the nation among the lowest of OECD countries.

Implementation Details

- **Data Preparation:** To prepare the data, in addition to standard pre-processing, we apply word lemmatization to remove prefixes and suffixes, as well as source-specific pre-processing using regular expressions(eg. removing HTML tags, etc.)
- **Training Details:**
  - Embedding: The Word2Vec model is trained on named entities to generate word vectors of **size 100**.
  - Clustering: For mini-batch k-means, we set the **batch size to 500** and tuned the number of clusters to obtain optimal k.

Cluster Tuning

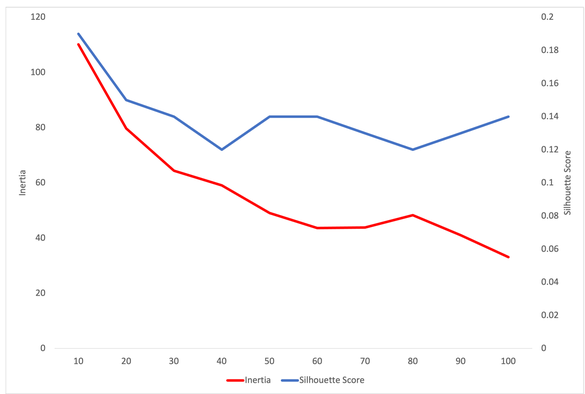


Figure: Silhouette Score/Inertia vs. Cluster Size

- As we increased the number of our clusters, both our inertia and silhouette coefficients decreased. While the goal of k-means is to minimize inertia, we chose silhouette coefficient to evaluate our clusters. Although a lower silhouette score indicates our clusters may have more overlap, **30 clusters** provided enough of a separation to get the newsworthy events while also keeping the clusters well-separated.

Future Work

- **Content Summarization:**
  - Produce salient and cohesive summaries from headlines or only the first/last paragraph.
- **Translation**
  - Create summaries in multiple languages to allow for breaking news updates globally.

Acknowledgements

- We would like to thank our mentors: Professor Kyunghyun Cho, Professor Hye Young You, and Inkeun Song for their guidance and support throughout the project. We would also like to thank our project advisor Najoung Kim for her constructive feedback and suggestions on many steps in the process.

References

[1] Salvatore Carta, Sergio Consoli, Luca Piras, Alessandro Sebastian Podda, and Diego Reforgiato Recupero.
Event detection in finance using hierarchical clustering algorithms on news and tweets.
PeerJ Computer Science, 7:e438, 2021.