Progress report

Angela Lam,Tingcong Liu

November 2023

1 Completed Tasks

Over the past month, we conducted research, find data sets, and wrote a pipeline for the model implementation.

Research (5H): while there are many research concentrated on differentiating major events,limited attention has been devoted to the task of key event detection. This task involves recognizing significant events within a set of documents centered around a common theme. In the initial phases of our project, we conducted research to explore methods for accomplishing this objective.

Data set (5H): The data set is obtained in the English section of the Miranda, Sebastião, et al. paper on Multilingual Clustering of Streaming News. Specifically, the data that will be used for this project is based on the 2014 Ebola outbreak with dates ranging from Sept. 18- Oct. 31 2024. The dataset contained articles that detailed the Ebola Outbreak. In addition, another file contained published dates of the documents. The date of publication of the document provides valuable temporal information that is useful in identifying peak phrases associated with specific events.

Model implementation (10H): So far, we built a pipeline detailing the structure for the code

- 1. vocab construction
- 2. TF-IDF
- 3. peak phrase detection
- 4. key event feature generation
- 5. document classification
- 6. post-processing

So far, we completed steps 1 and 2. This involved splitting documents into words, mapping documents with their dates, stop ward removal (using the Natural Language Toolkit), and eliminating rare words less than 10 occurrences. Finally the script complies a list of words that represents the initial vocabulary. Next, this data set-specific vocabulary is used in calculating TF-IDF scores.

2 Pending Tasks

Since we already constructed the pipeline, the pending tasks are implement steps 3-6. Step 3 involves generating candidate phrases from the vocabulary. Next, scores are calculated based on frequency and temporal distribution to determine peak words.

Step 4 identifies important time points and construct features based on the detected peak phrases in step 3. Documents are classified using this information (step 5). We also plan to do some post-classification processing (step 6).

After the code is constructed, we will do some experiments to optimize our results.

3 Challenges

One main challenge we faced is the granularity of the key events. Since all articles have the same overarching theme, they already shared the a lot of the same key phrases. It can be difficult to distinguish between different events due to the overlapping phrases and temporal overlap. Determining the appropriate level of granularity is crucial. If the granularity is too broad, the model will collapse the distinct events together, if it's too fine, our analysis might create too many insignificant event categories. It will require refinement of the current pipeline and tuning.

References

4 References

Miranda, Sebastião, et al. "Multilingual Clustering of Streaming News." arXiv.Org, 3 Sept. 2018, arxiv.org/abs/1809.00540.
Zhang, Yunyi, et al. "Unsupervised Key Event Detection from Massive Text Corpora - Arxiv.Org.", June 2022, arxiv.org/pdf/2206.04153.pdf.