The huge amount of news articles published everyday, particularly through online media outlets, poses a challenge to researchers in information retrieval: how to build a framework that automatically generates a timeline for news articles on a specific news story spanning days or even month. Such story comprises scores of salient events, each of which can be summarized by a textual event description. The corresponding timeline is a collection of these event descriptions, complete with time stamps on each update of the story. One of the most recurring problems suggested in the literature is how to best summarize an update/event. The problem consists two parts: (1) how to identify major updates, which are arranged in temporal order (2) how to select a sentence that best summarize each update. Some researchers have taken the approach of using news headlines or user-generated contents to assist with the event threading ([2], [3], [4], [5]). However, I am concerned with the quality and relevance of these types of supporting information. I am interested to investigate whether news articles alone would be sufficient to automatically generate timelines that present satisfying relevance and understandability. Specifically, many news articles includes the reporter’s commentary, which is not considered a part of the timeline and therefore should be separated at the very beginning.

I will use the corpus compiled by Wang et. al ([3]). The first step of the project is to separate factual statement (such as ) from opinion expression (such as ) at sentence level. I am planning to use Ozan Irsoy’s opinion mining system ([1]) for the task. Next is to identify the major events. Factual statements relevant to the identified event are extracted and ranked. Only one sentence is subsequently assigned to an update, forming a complete timeline. The automatically generated timelines will be compared to the gold-standard timelines provided by Wang et. al ([3]).
References


