# *Overview of various multi-document summarization methods*

Multi-document summarization is a technique that summarizes text from multiple documents automatically by extracting key pieces of the information, based on different heuristics or algorithms. This can help users with dealing with information overload that comes with learning about a specific topic that has a lot of documents. The goal is to produce an outline that sufficiently describes the topic in these documents/articles while being unbiased and comprehensive.

The aim of this technology review is to compare different techniques, especially in generating an outline from multiple news articles on the same topic. This would be extremely helpful in making a news topic summarizer/aggregator, as it deals with news "topics" or "themes". An example of a theme can be "Russian Invasion of Crimea", that can contain articles from New York Times and Russia Times. These two news sources would have different takes on the topic, but we want to just present the actual news of what's happening in Crimea, by converting both these documents into an objective outline.

Some of the characteristics that are more important in multi-document summarization than single-document summarization are compression, speed, redundancy, and passage. Since the algorithms will run over several documents at once, it has to scale better. A lot of times there are tradeoffs to consider such as analytical limitations, inaccurate extraction, essential sentences, low coverage, poor coherence and redundancy.

The basic goal is to create a compressed summary of the article set while retaining the main characteristics of the originals. Some methods use statistics to extract sentences from originals, producing a paragraph consisting of multiple sentences. Other methods analyze how the perception of an event changes over time, using multiple points of view over the same event or series of events. Yet others attempt to generate fluent text from sets of templates that contain the salient facts reported in the input texts.

After reviewing around 20 different techniques, the following 5 methods are the ones that were shortlisted to be reviewed and drawn conclusions from:

## 1. Timestamp Strategy based Naïve Bayesian Classifier [1]

This method uses a Naive Bayesian Classification and timestamp approach for summarizing multiple documents. The timestamp part achieves a coherent-looking summary and also provides it an ordered look. It works by extracting the most relevant part from the documents and using a scoring strategy to calculate the score for the words to obtain the word frequency. The higher linguistic quality can be estimated in terms of comprehensibility and readability.

## 2. Sentence-Level Semantic Analysis and Symmetric Matrix Factorization [2]

This method constructs the similarity matrix by calculating the sentence1-sentence2 similarities. It does this by doing some semantic analysis. The symmetric matrix factorization is

then used to group sentences into clusters. This has been shown to be equivalent to normalized spectral clustering. Finally, the most informative sentences are selected from each group to form the summary. It is a very fast and high performing method.

### 3. Cluster-Based Link Analysis [3]

The Conditional Markov Random Walk Cluster Model (ClusterCMRW) has, in the past, been used for summarizing documents under the assumption that all the sentences are indistinguishable from each other. It basically uses the link relationships between sentences in the topic. But for a given topic, especially pertaining to news, it usually covers only a few topic themes. These themes are represented by a cluster of sentences. The sentences in an important theme cluster are more "noticeable" by the method than the sentences in a trivial theme cluster and the topic themes are usually not treated equally. This method first uses link relationships to build topics and then uses two cluster models ClusterCMRW and ClusterHITS (Cluster-based HITS Model) to leverage the cluster-level information completely.

### 4. Simple Sentence and Metadata Extraction [4]

This method builds uses text extraction similar to ones used in single-document summarization methods, but adds additional information that is available to us about the whole theme and the relationships between the articles present in it. This can be done using many domain independent techniques based mainly on fast statistical processing. There is a metric used for reducing redundancy and maximizing diversity in the selected sentences. It can also specify a modular framework to allow easy parameterization for different topic characteristics and tweaks.

### 5. SUMMONS: A NLP Based approach to summarizing [5]

This method is good in summarizing a series of news articles based on the same time or event and is based on the traditional natural language processing (NLP) system architecture. It is a very modular system, meaning parts can be swapped in and out of the place. It is divided into two main components: a content planner and a linguistic component. The content planner selects information from an underlying knowledge base to include in a text, produces a conceptual representation of text meaning and does not include any linguistic information. The linguistic component, on the other hand, selects words to refer to concepts contained in the selected information and arranges those words. It then appropriately inflects them, to form a syntactically correct sentence. It does this by using a lexicon, which contains the vocabulary for the system and encodes constraints about when each word can be used.

### 6. Graph-based Neural Classifier [6]

This is a special type of machine learning method that basically uses sentence relation graphs. It employs a Graph Convolutional Network (GCN) on these relation graphs. It uses as input node features sentence embedding, which is obtained from a separate Recurrent Neural Network (RNN). The GCN generates high-level hidden sentence features, by going through multiple layer-wise propagations, for estimating prominence. It then uses various heuristics (swappable) to extract salient sentences. This classifier model avoids redundancy and improves by using

weighted features, compared to traditional approaches. It gains competitive results against other state-of-the-art multi-document summarizations but might be too resource intensive to incorporate in certain big projects.

## Evaluating Results and Conclusion

As we can see, above techniques can have different pros and cons, which may be highly subjective. For an objective approach of evaluating these methods, we have to put more emphasis on heuristics such as lower time, better precision, recall, f-score, scalability and accuracy.

Another way of evaluating these methods will be to generate a finite word summary (e.g. 100 words). Then take a look at the performance of most important cores of similarity measures, such as semantic similarity and sentence importance scores. Normalization may be used help to capture more information from the feature values.

I will be evaluating these throughout while working on the final project, but for now this overview should suffice.

## Works Cited


[1] N. Ramanujam and M. Kaliappan, "An Automatic Multidocument Text Summarization Approach Based on Naïve Bayesian Classifier Using Timestamp Strategy,"

[2] D. Wang, S. Zhu and C. Ding, "Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization

[3] X. Wan and J. Yang, "Multi-Document Summarization Using Cluster-Based Link Analysis,"

[4] J. Goldstein, V. Mittal, J. Carbonell and M. Kantrowitzt, "Multi-Document Summarization By Sentence Extraction,”

[5] K. McKeown and D. R. Radev, "Generating Summaries of Multiple News Articles,"

[6] M. Yasunaga, R. Zhang, K. Meelu, A. Pareek, K. Srinivasan and D. Radev, "Graph-based Neural Multi-Document Summarization,"