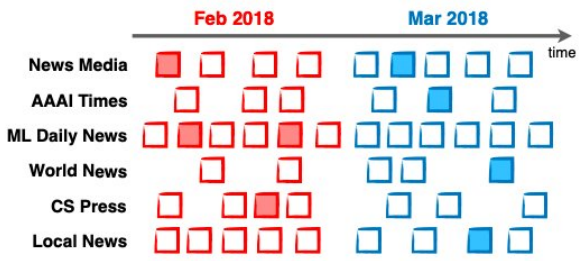
A new approach for comparative document summarization via classification

17 December 2018, by Ingrid Fadelli



An illustrative example of comparative summarisation. Squares are news articles, rows denote different news outlets, and the x-axis denotes time. The shaded articles are chosen to represent AI-related news during Feb and March 2018, respectively. They aim to summarise topics in each month and also highlight differences between the two months. Credit: Bista et al.

Researchers at the Australian National University (ANU) have recently carried out a study exploring extractive summarization in comparative settings. The term 'extractive summarization' defines the task of selecting a few highly representative articles from a large collection of documents.

In their paper, pre-published on arXiv and set to be presented at the 33rd AAAI Conference on artificial intelligence, the researchers considered comparative summarization, which entails the selection of documents from different document collections. These selected documents should be representative of each group, while also highlighting differences between the groups.

The project follows an ongoing theme at ANU's Computational Media Lab, which focuses on the automated understanding of large amounts of text and image streams on the social web. An overarching goal of the study is to identify techniques that could help people to deal with information overload.

"There is too much new content for anyone to read: news, social media feeds, or even the stream of arXiv research papers," Lexing Xie, one of the researchers who carried out the study, told TechXplore. "Can we ask computers to help us pick which one to read, and still receive crucial information?"

Xie and her colleagues have been investigating ways to summarize the hundreds of thousands of news articles, posts and discussions available online. Their aim is to present users with a few (e.g. 3-4) items that best answer the question 'what is new?' over a particular time frame (e.g. today, this week, etc.) or regarding a particular topic (e.g. climate change, elections, etc.).

"Text summarisation has been an active research field for almost 20 years, but the main focus has been to summarise one collection either extractively (i.e. select existing items to compose a summary), or abstractively (i.e. composing new sentences as summary, rather than using existing ones)," Xie explained. "This work focuses on extractive comparison of document groups, i.e. selecting a few items from a group that is most distinct from other groups. To the best of our knowledge, our work is the first to carry out and validate comparative summarisation at scale."

In their study, the researchers approached comparative document summarisation as a classification task. Classification is a common machine learning task, in which an algorithm makes educated guesses about what category or groups particular data items belong in.

"In the case of comparative summarisation, if we have chosen good summary articles it should be difficult, if not impossible, to design a classifier that can distinguish between the chosen summary articles and the groups to which they belong; while it should be easy to design a classifier that can distinguish between the chosen summary articles

and other groups," Alexander Mathews, another researcher involved in the study, told TechXplore.

The classification perspective taken by the researchers entails an alternative but complementary view of comparative summarisation as three competing objectives. First, selected summary articles should be representative of the groups to which they belong, covering all important aspects of the document collection.

Second, each chosen summary article should be relatively different from the others, in order to avoid unnecessary repetition. Finally, selected summary articles should only be representative of the group to which they belong, as this is a key factor for effective comparative summarisation.

"Our specific formulation of the three objectives relies on a flexible mathematical measure called the Maximum Mean Discrepancy (MMD)," Mathews explained. "This measure, along with the application of a mathematical tool called 'the kernel trick' allows us to cast our three objectives into a compact mathematical form which we can optimise efficiently even on huge datasets. Moreover, this form permits both discrete and gradient based optimisation techniques, allowing the choice of articles to be finely tuned to meet our objectives."

The classification perspective taken by Mathews and his colleagues allowed them to evaluate their method as a classification task, both automatically and via crowdsourcing. Their approach outperformed discrete and baseline approaches in 15 out of 24 automatic evaluation settings. In crowdsourcing evaluations, summaries selected using their simple gradient-based optimisation strategy elicited 7% more accurate classification from human workers than discrete optimisation methods.

"We are glad to see that using only 4 summary articles per week the accuracy of automatic classification (of each news article into the month/week that it came from) is on par with one that 'reads' all articles," Minjeong Shin, one of the researchers who carried out the study, told TechXplore. "This demonstrates that crucial new information is contained in the few 'prototype' articles."

The researchers evaluated their method against other approaches on a newly curated collection of controversial news topics spanning over 13 months. When applied to the comparative summarisation of ongoing content streams, their system successfully answered questions such as 'what is new on the topic of climate change this month?', highlighting differences between two distinct time periods.

"Our methodology also applies to collection comparisons other than news over time," Shin said. "For example, one can ask: what is the difference between BBC and CNN coverage of the G20 summit, or how does the coverage of climate change differ between UK and Australian media?"

In the future, this new approach to comparative summarisation could help users to navigate the large amounts of information available online; providing comparisons of articles published by different sources or authors, as well as of posts on related topics or expressing distinct viewpoints. The researchers are now working on expanding their research by taking these comparisons to the next level.

"We are investigating ways to summarise not just text, but also images and text jointly," Umanga Bista, one of the researchers who carried out the study, told TechXplore. "We would also like to take into account known relationships of entities mentioned in the text (e.g. Delhi is the capital of India), rather than treating each word as an independent entity. Ultimately, we would like to have a system that recommends what is new, what is different, and what is worth reading."

More information: Comparative document summarisation via classification. arXiv:1812.02171 [cs.IR]. arxiv.org/abs/1812.02171