Title

Citation Spans of Hyperlinks in Change My View Posts

Abstract

"Change My View" (CMV) is a forum on the popular website Reddit where users are invited to post interesting opinions and challenge others to change their minds. Each thread is started by an original post (OP) that includes a title of the format "CMV: *Original Poster's opinion here*" and text explaining this viewpoint. The OP is followed by a series of comments arguing against the Original Poster's position, and particularly convincing cases are awarded by a delta (Δ). These characteristics of CMV posts make this Subreddit a popular source of casual persuasive text.

Within Change My View OPs and comments, which we will collectively refer to as posts, we focus on the presence of hyperlinks. When presenting arguments for their opinions, users sometimes provide evidence to back up their claims in the form of a hyperlink to a separate source; the hyperlink is often embedded within a phrase without disrupting the flow of information. During this study, we observe the citation span of the hyperlinks and explore the possibility of text surrounding a hyperlink serving as a summary of the content of the external source.

After manually annotating each post for the predicted value of the citation spans, we set a window for the automatic labeling of citation spans (the sentence before the sentence containing the hyperlinked content). Over a thousand posts were collected; 7.0% of these posts contained hyperlinks with an average prevalence of 1.42 links per post with hyperlinks. This shows that while a small fraction of users tends to include hyperlinks in their posts, those who do tend to include more than one. Most sources and articles come from news websites like the New York Times, Washington Post, and Huffington Post, and a significant number of articles refer to YouTube and Wikipedia. Along with basic characteristics of the dataset, the precision and recall were also recorded for each post. The trend shows high precision with varying recall, which indicates that the automatic window was successful in many cases but not all.

This study was designed as the first step for further investigation into citation spans, summarization, and argument generation. It successfully provides a systematic way to characterize the dataset of hyperlinks and a program to label and analyze citation spans against predicted values.