Extracting narratives from Twitter data in real time

Oftentimes, small- and big-time events are discussed in microblogs (usually associated with social media, such as Twitter) before they become more formal news articles. Artificial Intelligence can help to transform scattered large amounts of microblogs (for example, tweets) and uncover a story about the event behind them. However, given the short nature of tweets, they are usually written in more colloquial language (contrary to the formality of news articles), which proposes an interesting challenge NLP-wise, since most NLP tools are only prepared to deal with more formal language.

In this project, the use of a combination of NLP techniques is proposed, such as named entity extraction (NER), coreference resolution (COREF) and semantic role labeling (SRL) to extract the relevant narrative elements from tweets.

There are two main goals in this project, the first is to summarize a set of tweets (about a topic) and understand if they can effectively summarize a news article (about the same topic). The second is to automatically represent a set of tweets about a topic through a formal narrative representation called Discourse Representation Scheme and complement it with a Knowledge Graph.

To this extent, the development and evaluation set used is the *Signal-1M tweetir* dataset, which links several tweets about a certain topic with a news article about the same topic. This dataset is used to test our NLP approaches and to understand to what extent the content of the tweets is representative of the content of the news article.

Finally, as for results, some baseline summarization approaches were tested, such as LexRank, LSA and even a pre-trained BART model and they all had poor performances, around 20% similarity (measured by ROUGE) between tweets and news about the same topic. However, using the proposed methods (COREF and SRL) similarities of over 50% between main parts of the tweets and the article were obtained.