April 29 2021

Extracting Text into Meta-Data

Improving machine text-understanding of news-media articles

Author licentiate thesis: **Johannes Lindén**
STC Research Centre
Mid Sweden University

Abstract

Society is constantly in need of information. It is important to consume event-based information of what is happening around us as well as facts and knowledge. As society grows, the amount of information to consume grows with it. This thesis demonstrates one way to extract and represent knowledge from text in a machine-readable way for news media articles. Three objectives are considered when developing a machine learning system to retrieve categories, entities, relations and other meta-data from text paragraphs. The first is to sort the terminology by topic; this makes it easier for machine learning algorithms to understand the text and the unique words used. The second objective is to construct a service for use in production, where scalability and performance are evaluated. Features are implemented to iteratively improve the model predictions, and several versions are run at the same time to, for example, compare them in an A/B test. The third objective is to further extract the gist of what is expressed in the text. The gist is extracted in the form of triples by connecting two related entities using a combination of natural language processing algorithms.

The research presents a comparison between five different auto categorization algorithms, and an evaluation of their hyperparameters and how they would perform under the pressure of thousands of big, concurrent predictions. The aim is to build an auto-categorization system that can be used in the news media industry to help writers and journalists focus more on the story rather than filling in meta-data for each article. The best-performing algorithm is a Bidirectional Long-Short-Term-Memory neural network. Three different information extraction algorithms for extracting the gist of paragraphs are also compared. The proposed information extraction algorithm supports extracting information from texts in multiple languages with competitive accuracy compared with the state-of-the-art OpenIE and MinIE algorithms that can extract information in a single language. The use of the multi-linguistic models helps local-news media to write articles in different languages as a help to integrate immigrants into the society.

Sensible Things that Communicate
PART OF MID SWEDEN UNIVERSITY