Deep Learning for Natural Language Understanding and Summarization

Candidate: Moreno La Quatra
Supervisor: Luca Cagliero

Politecnico di Torino

Abstract

Information overload is a major problem affecting today's society. The continuous stream of information makes it impossible for individuals to process and understand the vast amount of information available to them. Automatic content understanding is a research field that aims to alleviate this problem by efficiently extracting relevant information from unstructured sources (e.g., text documents). Natural Language Processing (NLP) is a branch of artificial intelligence involving automatic interpretation and manipulation of human language. The use of NLP techniques is key to many automatic content understanding tasks, such as information extraction and text summarization.

Our research aims at investigating and developing methods for automatic content understanding to alleviate the effects of information overload. We focus on three specific data types: academic articles, news stories, and podcasts. In order to automate the extraction of relevant information from unstructured sources, we explore and present several methodologies tailored to each domain.

Our contributions in the field of scientific literature analysis aim at providing researchers with resources to navigate the ever-growing amount of scientific papers. First, we analyze citation context to assess whether reading the full text of a publication is likely to be useful for understanding the referencing paper, thereby reducing the amount of time spent on literature search. Second, we devise two different models that process scientific papers to extract a results-oriented overview of its main contributions (namely, the highlights) and facet-specific summaries tailored to the needs of different types of readers. Finally, we explore the potential of unsupervised summarization models to generate slides for scientific talks, thus providing authors with an initial draft of their presentation.

As part of our work in the news domain, we concentrate on the task of news timeline generation, a process for automatically creating a timeline of specific events that summarizes the most important developments that happened during event unfolding. We present a pipeline that summarizes what happened at each date and then selects salient dates according to graph-based centrality measures. Since the methodology is specifically tai-

lored to the English language, we separately investigate the same problem in multilingual settings and propose a solution that, given a set of documents in multiple languages, incorporates cross-lingual alignment and automatic machine translation to generate timelines in a target language.

Finally, we investigate the challenge of podcast summarization and propose an extractive summarization approach that leverages multimodal information (i.e., audio and text) to identify the most important segments of an episode. Our system relies on an end-to-end deep learning architecture that jointly learns feature representations from both modalities. It combines text and audio features using a late-fusion strategy and is trained to score and select relevant segments from the spoken content. By automatically selecting and concatenating the most important portions, this methodology can provide both a textual and an audio summary of a podcast episode. Our system is further enhanced by proposing a select-and-rewrite approach that can generate a more fluent summary by rearranging and rewriting selected segments using a state-of-the-art sequence-to-sequence model.

The research domains investigated in this thesis share the need to manipulate linguistic information and, more specifically, human language. Therefore, we extensively use NLP techniques to understand the semantics of language and automatically process unstructured content. Considering a broad range of scenarios, this thesis examines how NLP can enhance the understanding of human language and facilitate the processing of large volumes of data.