Abstract

The exponential growth of information accessible on the World Wide Web has posed a significant challenge in processing and summarizing vast textual content. This has led to the development of automatic text summarization systems to condense lengthy documents into manageable forms while retaining crucial information. Our research endeavors to tackle various challenges in automatic text summarization, focusing on both resource and model development.

Recognizing the unique nature of tables as a form of data representation, we embarked on exploring table summarization systems. However, table summarization faces a scarcity of large and diverse datasets for effective model training, unlike other natural language processing tasks such as machine translation or sentiment analysis. To overcome this hurdle, we delved into methods for crafting structured datasets tailored specifically for summarizing tables, thereby facilitating the development of robust summarization systems. Our work introduces several iterations of gold standard table summarization datasets. Additionally, we proposed the performance analysis dataset, further enriching our resources for generating student performance summaries. These datasets undergo meticulous annotation to ensure they provide relevant information for training effective summarization models.

In the domain of extractive summarization, we put forth novel methods, including rule-based and template-based approaches, to effectively summarize tabular data. These methodologies enhance the quality and relevance of extracted information, preserving essential content within the summaries. Following this, we developed a sense-based model aimed at effectively ranking and summarizing English queries to align with the objectives of the NTCIR shared task. Leveraging sentiment lexicons and tabulation-based methodologies, our system yielded positive outcomes, demonstrating its ability to reduce user interaction with mobile devices while extracting relevant information.

Furthermore, we conducted a study on summarizing scientific articles, as evidenced by our participation in the SCISUMM 2017 shared task. Drawing on our expertise in

extractive summarization, we devised methods to identify significant information within scientific literature. These endeavors contributed to enhancing knowledge dissemination and facilitating information retrieval within academic communities.

Subsequently, we explored extractive summarization using word embedding techniques, aiming to leverage word vector semantic representations to enhance the coherence and informativeness of extracted summaries. These models were trained and tested on open source gold standard datasets like CNN/Daily mail.

In the realm of abstractive summarization, we leverage deep learning methods such as LSTM and RNN architectures to craft coherent summaries. Through extensive training and testing on reliable open-source datasets, we ensure the accuracy and relevance of our abstractive summarization models. Moreover, we introduce a novel strategy using one-word abstractive summaries derived from our proposed performance analysis datasets, offering concise insights into student achievement.

A significant contribution of this thesis also lies in proposing and evaluating innovative table-based summarization systems that leverage extractive summaries to generate abstractive summaries. Noteworthy among these models are the "Fine-Tuned T5" model, a transformer-based approach fine-tuned on our custom dataset, a sequence-to-sequence encoder-decoder model, and the "SETA" model, an attention-based encoder-decoder model focused on similarity. These models undergo rigorous evaluation using standard metrics, showcasing their ability to produce coherent and informative summaries from table-based content.

Moreover, our research extended to the field of ranking in natural language processing, where we proposed novel models for query ranking tasks. By participating in collaborative tasks such as the NTCIR-12 shared task and the IJCNLP-17 Review Opinion Diversification task, we developed sense-based ranking systems and feature classifier based models for ranking queries.