22/10/2019

Generating Hotel Multi-Review Summarizations

Description:

Nowadays, information from the Web is overwhelming us from all sides and creates a need for automated multi- document summarization systems that produce high quality summaries. Extractive multi-document summarization, where the final summary is composed of sentences in the input documents, has been addressed by a large range of approaches. Generally, summarization systems output summaries in two steps: sentence ranking followed by sentence selection. The first estimates an importance score of each sentence and the second chooses which ones to select by considering 1) their importance and 2) their redundancy among other sentences and the current summary. Most of the time, all sentences in the same collection of documents are processed independently and therefore, their relationships are lost.

TripAdvisor[1] is a famous website showcasing hotels with reviews. Users can write freely about their experiences with various hotels and also assign sub-ratings to specific aspects, such as cleanliness, location, or service. However, the number of hotel reviews can quickly be overwhelming, making it difficult for a user to find the information (s)he is looking for.

The idea of this project is to study **unsupervised** multi-reviews summarization; as gold standard are not available, we will be working on different variations of the new joint work of Google and MIT [1]. The proposed model is based on auto-encoders that first encode k reviews in a shared embedding space and decode them to reconstruct the original reviews. In parallel, the model aggregates the review representations into a single embedding and decodes it to generate the final summary.

The idea of this project would be to experiment with different aggregation functions with parametric and non-parametric models: mean, median, feedforward neural network, recurrent neural network, etc.

As a first step, we would verify the code available in [2] by running their method on a restaurant dataset and then apply the method on hotel reviews. In a second step, we would modify the code to include different aggregation functions.

Prerequisites: Knowledge about Machine Learning, Tensorflow and efficient with Python.

Supervisor: Diego Antognini (diego.antognini@epfl.ch)

References:

[1] https://arxiv.org/pdf/1810.05739.pdf
[2] https://github.com/sosuperic/MeanSum

[1] https://www.tripadvisor.com/