NLP Project Proposal - Robin Cosbey and Josh Loehr

1. What is the problem or task you propose to solve?

We propose the summarization of scientific papers by way of machine learning methods.

2. What is interesting about this problem from an NLP perspective?

Research at the university level requires a large amount of literature review and the number of relevant scientific papers to a project is always growing. Having the ability to generate a summary of a paper's contributions would make this process more efficient and increase a researcher's ability to understand the scope of their project early on.

3. What technical method or approach will you use?

We plan to approach the automatic summarization task with extractive machine learning binary classification methods in which sentences of provided documents are labeled as either *summary* or *non-summary* (Saranyamol et al., 7890). Relevant sentences will be compiled in chronological order to form a summary of the document. This will be accomplished by way of neural network architectures (Das, 7). We will have two baselines to compare our results against. The first will consist of the first sentence from each paragraph in a document. The second will consist of the highest TF-IDF scoring sentence from each paragraph of the document. We are considering several feature variants which include sentence embeddings, frequency of terms, the presence of title words, and sentence length (Kumar et al., 180).

4. On what data will you run your system?

TIPSTER SUMMAC: 183 scientific papers from Association for Computational Linguistics (ACL) sponsored conferences in xml form with abstracts for comparison http://www-nlpir.nist.gov/related_projects/tipster_summac/cmp_lg.html

MEDLINE: 1,875,206 biomedical literature articles in xml form with abstracts for comparison https://www.nlm.nih.gov/bsd/pmresources.html

5. How will you evaluate the performance of your system?

To initially assess the performance of our model, we will evaluate the generated summaries manually. We will also implement ROUGE metrics to determine the quality of a model-generated summary by comparing it with the provided document abstracts (Lin, 1).

6. What NLP-related difficulties and challenges do you anticipate?

Given that we are approaching the text summarization task with machine learning methods, collecting and processing documents will be a primary challenge - we will need our documents to be uniform in terms of formatting and feature structure to produce consistent classifications. We may also encounter issues with out-of-vocabulary sentences; we will need to employ methods to address this without losing accuracy. A third challenge may be the generalizability of our model. If we choose to train on documents from a single domain, the trained model may not perform well on documents from other domains.

References

Das, D. and Martins, A. “A Survey on Automatic Text Summarization.” Literature Survey for the Language and Statistics II course at CMU, 2007, pp. 192–195.

Kumar, Y. et al. “A Review on Automatic Text Summarization Approaches.” Journal of Computer Science, vol. 12 (4), 2016, pp. 178–190., doi:10.3844/jcssp.2016.178.190.

Lin, C-Y. “ROUGE: a package for automatic evaluation of summaries.” In Proc. of the ACL 2004 Workshop on Text Summarization Branches Out, 2004, pp. 74-81.

Saranyamol, C. et al. “A Survey on Automatic Text Summarization.” International Journal of Computer Science and Information Technologies, vol. 5 (6) , 2014, pp. 7889-7893.