UNIVERSITY OF PITTSBURGH, SCHOOL OF COMPUTING AND INFORMATION

INFSCI 3005 BLOG

Anthony Sicilia

2019-09-12
1) Talk Summary
Duolingo.AI Panel Discussion
I attened a pannel discussion with a group from the machine learning and data science teams at Duolingo. There were 2 machine learning engineers – one of whom was actually working in computer vision rather than NLP – a research scientist, a data scientist, and the research director who coordinated the discussion. Discussion was primarily on the different types of projects on which each member was working as well as some commentary on what it is like to work at Duolingo vs. other companies/academia.

The first machine learning engineer (in computer vision) described how Duolingo uses computer cameras to identify whether test takers are cheating when they use the Duolingo proficiency exam. The task is basically anomaly detection with a lot of difficult challenges because almost no assumptions can be made on the environment in which the test will be taken. The second machine learning engineer worked on predicting user language proficiency from text as well as a smart tips feature. Both of these two engineers discussed the importance of human experts and human-in-the-loop AI. The research scientist was a professor at Northwestern who now works on student modelling problems for Duolingo. For example, one use case of his work might be understanding how a specific user will respond to a newly proposed language challenge. The data science member was the last to describe his role and he was very clear on the difference between machine learning and data science from the perspective of Duolingo. Basically, where machine learning engineers will implement algorithms whose outputs become actual feature in the app, the data science teams spend time forecasting, visualizing data, and in general trying to communicate information so that other Duolingo teams (for example, the marketing team) can make informed decisions. Overall, the panel was informative on the types of work and research that go on at Duolingo.

2019-09-05

1) Talk Summary

Improving Sentence Retrieval from Case Law Statutory Interpretation presented by Jaromir Savelka

Summary: This talk talk discussed the creation of a dataset as well as the proposal of a new methodology in the realm of passage retrieval. The dataset and methodology are specific to sentence retrieval in case law where the goal is to assist the end-user in acquiring a ranked list of passages for the purpose of interpreting certain terms specified in a law. For example, a law specifying the prohibition of cars in a certain location might use the term "vehicle." In this context, the ranked list of passages would assist the end-user in interpreting the term "vehicle" and making an argument for how this term should be used in a court of law.

One contribution listed in the talk is the creation of a dataset to train machine learning methods. This dataset consists of annotated passages containing a few selected terms. Here, the annotation (by a human expert) denotes the quality of the passage; i.e., it establishes a ground truth for the ranking process.

A second contribution is the proposed methodology. The authors tested a number of methods including specialized tf-idf methods and similarity comparisons between word embeddings, for example. They explore a number of additional considerations in the retrieval process including context, query-expansion, novelty detection, and ultimately an agglomerate of all three of these. Comparison of all of these methods with a variety of implementations is provided.

Speaking with the presenter after the talk, one thing I was interested in was how word-vector based methods were aggregated (for comparison of entire sentences). An average was used which could explain the poor performance of *dense* word-vector based methods on longer sentences. It seems reasonable that the longer a sentence is the less representative this average of dense word-vectors would be (especially, if on short sub-passage within the parent passage is the most important to consider). With this said, it would be interesting to consider neural encoder-decoder methods with attention to get a passage representation; these methods may be better able to highlight pertinent information in the representation (as they have done in other natural language tasks).

Acknowledgements

LaTeX template by Phillip Mak under a Creative Commons License