Project Individual 5: UIMA Metrics

Leah Nicolich-Henkin: lnicolic

October 5, 2015

Overview

This report outlines a UIMA architecture to calculate metrics for a system that ranks different passages based on whether they provide answers to a given question. The system is nearly identical to that presented in PI4, so I won't go into the details, but rather will focus on the differences.

Type System

As presented in PI4, my type system was changed to add the `QuestionSet` type. Each `QuestionSet` includes a `Question` and an FSArray of `Passages`. This makes it easy to iterate through all the annotations and have everything that shares an ID conveniently associated.

The only change to the type system I made in this project was to add metric features. The `QuestionSet` now has the features *pAt1*, *pAt5*, *rr*, and *ap*. This makes it easy to calculate the metrics as part of an AAE, and then reference them from the consumer.

Annotations

There are five annotators: Question Annotator (annotates questions), Passage Annotator (annotates passages), QuestionSet Annotator (combines questions and passages), Score Annotator (calculates a score for each passage), and Metric Annotator (calculates metrics). The only change between PI4 and PI5 is the addition of the metric annotator.

Metric Annotator The Metric Annotator iterates over the Question Sets, calculating the appropriate metrics for each one. For each Question Set, it iterates over the passages, adding up the total number of correct responses, calculating precision at each cutoff, and setting the reciprocal rank the first time it encounters a correct answer. After iterating over all the Question Sets, it is able to calculate Mean Reciprocal Rank and Mean Average Precision for the entire set.

Consumer

The consumer is the other component that changed from PI4. Instead of writing the scores of the passages, it writes a csv file containing each of the calculated metrics.

Metrics

The four metrics presented are precision at 1 (P@1), precision at 5 (P@5), mean reciprocal ranking (MRR), and mean average precision (MAP). P@1 serves as a good simulation of what would happen if the system were being used, and was actually returning a single passage to each question. It's a straightforward "yes, you got it right", or "no, you didn't". However, it doesn't provide much nuance. If it's wrong, there's no feedback on how far off it was, and if it's right, is it just luck, or are all the passages actually ranked correctly? In addition, there might be times when you want more than one correct passage, for example to compare answers. P@5 offers a solution to some of these problems by providing feedback on a few ranked passages, instead of just one. This means that it values ranking, rather than just getting the top one right. One downside, however, is that in cases where there are less than five correct passages, it will necessarily score low, even if the ranking is correct, thus unfairly bringing down the overall score.

MRR is similar to P@1 in that it only cares about one correct answer, presumably with the assumption that you're only returning one correct answer. However, it also gives "credit" for not getting the top-ranked passage correct, with a higher score the closer a correct passage is to the top of the ranking. This metric takes into account the idea that in the end you're only returning one passage, but also can show improvement as other non-top-ranked passage improve in ranking. Finally, MAP gives the most wholistic measurement of everything going on in the passage ranking, taking into account not only precision but also recall, and using a variable N that goes through all possibilities.

Performance

Currently, the ranking method implemented in the Score Annotator is the MITRE method of word overlap, which is essentially a unigram comparison. The score is the percentage of words in the question that also occur in the passage. It is not a particularly powerful method, but serves as a good baseline for future work.

Calculating metrics makes its bad performance quantifiable. The MRR is **0.422**, and the MAP is **0.393**. Having these metrics is a powerful tool, because as I start to make improvements in the next few weeks, I will be able to judge whether they are meaningful improvements based on calculations rather than intuition or guesswork.