

Homework 2  
11-791  
Fall 2013

Callie Vaughn (cvaughn)

I have created a system that judges the similarity between a question and multiple answers using N-Gram similarity, then evaluates itself by a precision at N measure.

The system first extracts tokens from the text. These tokens include the “Q” and “A” labels as well as the “0” and “1” values noting whether each answer is correct. Next, the system takes all tokens other than those described above and groups them into all possible unigrams, bigrams, and trigrams for each sentence. The questions and answers are then labeled distinctly from each other. Then the N-Grams contained within the span of the question (i.e. those N-Grams that make up the question) are compared to the N-Grams of each answer in turn. These comparisons are used to calculate a score for each question as follows: the count of N-Grams that are in both the question and answer is divided by the total number of N-Grams in the answer.

After the scoring as described above is complete, the answers are sorted in order of their score, from high to low. Because we have a gold label telling us whether each answer is really correct or not, we know the number of correct answers that exist. Call this number N. We take the N highest-scoring answers and only consider them. Of those N answers, those with scores higher than the threshold (0.3) are judged as positive by the system, while those with scores below the threshold are considered negative. The number of true positives (those correctly identified as positive by the system) is divided by the number of true positives plus the number of false positives (the total number of answers judged as positive by the system) to yield the precision at N of the system.

For the first example (“Booth shot Lincoln?”), the system has two true positives and two false positives, leading to a precision of  $2/4 = 0.5$ . For the second example document (“John loves Mary?”) there are two true positives and zero false positives, leading to a precision of  $2/2 = 1.0$ .