Basic System - Reading Comprehension on Short Passages for Question Answering

JAKE PITKIN
CS 6390 - Information Extraction
April 15, 2017

Evaluation Results

Below is a table of the accuracy and F1-score for various systems and benchmarks. To evaluate my system, I used only one of the sixty categories of documents and Q/A's for time reasons (would take multiple days train a system across all the training examples).

Approach	Dataset	Accuracy	F1-Score
Random Guess	Dev	1.1%	4.1%
My Basic System	Dev	5.8%	10.7%
My Final System	**Dev**	**14.57%**	**21.64%**
State-of-the-art	Test	76.92%	84%
Human Baseline	Dev	80.3%	90.5%

Table 1: Bold indicating my current system. Approaches in ascending order based on performance.

Additionally I wanted to evaluate how well the heuristic of using the noun phrases in the document as the candidate pool of answers for a question. I took 2,000 questions and computed how often one of the possible answers to a question ended up in the candidate pool.

Questions	Answers Found	Percent
2,000	1,105	55.25%

This approach puts a low ceiling on the possible recall of my system. A better approach to finding candidates would be worth exploring.