Applying Network Centrality Techniques to Question Answering

Pedro Rodriguez Fenfei Guo

October 5, 2016

1 Project Proposal

Question answering is an important domain which traditionally combines machine learning, information retrieval, and natural language processing techniques to create intelligent agents capable of answering human-posed questions. In this project we limit our focus to trivia and factoid questions. We enforce the following constraints: 1) questions must have a single unique answer, 2) each sentence in the question text must uniquely identify the answer, and 3) each answer must correspond to a single Wikipedia page. All of these requirements are met by a particular question answering task based on a popular trivia game called Quiz Bowl [1]. In the task, questions such as the one in figure 1 are read by an arbitrator to two competitors. As soon as any competitor knows the answer they are allowed to "buzz in" and provide an answer. If their answer is correct they are awarded points, otherwise they are penalized and the other competitor has a chance to respond.

Formally this problem can be framed as multi-class classification over the set of possible Wikipedia entries. Prior approaches have further constrained the answer set by only using classes which occur in the Quiz Bowl dataset, and additionally for the purposes of making deep learning methods more effective only using classes for which there are at least a minimum threshold of training examples [3]. The goal of this project is to devise a method which drops at least the second constraint which would bring the answerable number of classes from approximately 2,000 to approximately 10,000.

To accomplish this we would first construct the full Wikipedia graph where vertexes are entries and edges indicate the existence of a hyperlink between one page and another. The second step of the process is to label all the named entities in the question text read so far with their corresponding Wikipedia entry. This step is already partially implemented in the Qanta AI project [5] using the Illinois Wikifier [2, 4]. The output of the Wikifier is shown in figure 1 as blue words. Shown in red are a subset of entities we believe would be easy to map to Wikipedia entries by exact matching of unigrams and bigrams to wikipedia page titles. Next we propose constructing a subgraph containing the extracted Wikipedia entries and any nodes which are nearby to each of these vertexes (eg distance 1 or 2 away). In the construction of this subgraph the weights could be weighted according to features such as TF-IDF cosine similarity. Intuitively we would expect that the vertex corresponding to the answer (a Wikipedia entry) should included in this subgraph and have higher centrality than non-answer vertexes. Our primary research goal in this project is to identify or create vertex-level centrality metrics which are predictive of the true answer class which in the graph is represented as a vertex.

Figure 1: Sample quiz bowl question whose answer is Albert Einstein. Wikipedia entries linking to Albert Einstein's entry are bolded, blue entries indicate Wikifier output, and red entries correspond to some unigram/bigram title matches.

With **Leo Szilard**, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity*, so **Debye** adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in **tensor** products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking **1905** papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this **German physicist** best known for formulating the special and general theories of relativity?

References

[1] Jordan L Boyd-graber, Brianna Satinoff, He He, and Hal Daumé III. Besting the Quiz Master: Crowdsourcing Incremental Classification Games. *EMNLP-CoNLL*, pages 1290–1301, 2012.

[2] X. Cheng and D. Roth. Relational inference for wikification. In *EMNLP*, 2013.

[3] Mohit Iyyer, Jordan L Boyd-graber, Leonardo Max Batista Claudino, Richard Socher, and Hal Daumé III. A Neural Network for Factoid Question Answering over Paragraphs. *EMNLP*, pages 633–644, 2014.

[4] L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In *ACL*, 2011.

[5] Pedro Rodriguez, Jordan L Boyd-graber, and Mohit Iyyer. Qanta: Quiz bowl ai. `https://github.com/Pinafore/qb`, 2016.