Initial Report

1. Team Members

Rainer Nunez (rnunez), Zhenzhen Weng (wengz), Xi Liu (xiliu1), Yiting Huang (yitingh)

2. Use of Development Data

Development data includes the Wikipedia articles and the questions/answers that we generated. For answering, we will use it as a golden standard to examine the performance. For questioning, we will compare questions generated by the system with questions generated by ourselves and improve our model. Besides, if newly generated questions are good enough, we will add them and their answers to our development set.

3. Relationship between Q&A

The question system will rely on the answering system because we believe if a question could get good quality, it is more likely to be a good question. Thus we will develop the answering system first and then based on the ability of getting good answers, we then develop the questioning system.

4. Toolkits

Python NLTK, Berkeley Parser, Porter/Lancaster Stemmer, Stanford NER Tagger, Supersense, QuestionBank, WordNet

5. Collaborations

We will share code and data on Github
Org: https://github.com/NLPQA
Repo: https://github.com/NLPQA/WikiQA

We will communicate through GroupMe, Google Drive and emails.

6. Technical Approaches

All four will develop the answering system first and then based on the ability of getting good answers, we then develop the questioning system. After discussion and reading, we have a general idea of developing two systems.

6.1 Data Preprocessing

For both questions and documents, we need to tokenize and stem them, also remove stop-words. Name entity recognition could be applied to to identify what to ask or choose the most reasonable answer.

6.2 Answering

There are two types of questions to answer, yes/no questions and W/H questions. We need to think about different approaches.

For yes/no questions, we will convert the questions into a statement and try to retrieve the document with highest tf-idf score. Then calculate the cosine similarity to select the sentence and check the fact with it, then return yes or no.

For W/H questions, before sentence selection, we need to identify the purpose of the question, and find most similar sentences with the target, which could be classified by NER or identified by some keywords.

To evaluate the result, we will utilize the development data. For yes/no questions, we could calculate MAP for each run while for W/H questions, compute the similarity of the answer with gold standard. In addition, we will manually inspect results.

6.3 Questioning

We only use sentences which are long and informational enough to generate questions. Currently we plan to select those sentences and represent them as parse trees or vectors with keywords only. Then we substitute bodies with who/what, belongingness with whose, choices with which, time or locational information with when/where, and reasons with why.

To measure the quality of questions, we plan to check its grammar and ability to get good answers, which relies on the answering system. We will manually inspect results, too.