11-411 Project Initial Report

Mingquan Chen, Yuzhe Sun, Qiyang Xu, Juncheng Zhan

February 13, 2017

Abstract

This is the initial report for team Chuxi's Natural Language Processing project. In this report, we will discuss rough ideas on the design, development, and evaluation of the question answering system, especially the tools and techniques. We will also discuss how to share resources and coordinate among team members.

1 Use of development data

The development data will be used to build the asking and answering components for the 10 known classes of Wikiepedia pages; i.e., the model already captures information from the development data, possibly in the form of a database. For the "secret" class of Wikipedia pages, both the asking and the answering components will be built agnostic on the development data, and the data will be used to evaluate the performance of the models; finally, model selection will be performed on these models manually or by cross validation using ten given categories.

2 Relationship Between Question Generation and Answering

The asking and answering components **will** be related; the rationale behind this is that they share the same information derived from the passages. The passages will be processed to establish a database from which both the asking and answering systems can retrieve information. Further, the answering system will be used to evaluate the effectiveness and difficulty of the questions produced by the asking component.

3 Communicatation and Coordination Inside the Team

3.1 Sharing Method

We will use git (and a GitHub repository) for data and code sharing. Documentation will be written collaboratively in Overleaf.

3.2 Coordination of Development

The team will discuss and come up with general approaches together, and then split up implementation of algorithms to individuals. As the project initiates, communication is the key. Therefore, we plan to talk about project whenever we meet in class and have a weekly meeting to formally discuss our progress and decide on the direction. Facebook/Wechat will be our communication platforms outside class. Teammates will commit/push to Github and post announcements whenever they have new ideas or finish certain parts. Juncheng Zhan and Yuzhe Sun are responsible for database construction; Qiyang Xu and Mingquan Chen need to generate asking and answering components based on the database. For the "secret" case where we don't know the category of the passages, we are going to discuss further and split work to each person as the work progresses.

4 Tools

We are going to use Python as the main programming language. Within Python, we plan to use the regular expression library, as well as the string library. In addition, we might use TensorFlow to assist us with deep learning aspect of NLP, and spaPy for general NLP tasks.

5 Technical Approaches

First, we are going to employ several machine learning techniques for this project. These techniques include Decision Tree, Neural Networks, Deep Learning, Hidden Markov Models, Bayesian learning, etc. In addition, we will make use of the above machine learning techniques to implement algorithms for named-entity recognition, parsing for grammatical analysis, relationship extraction, text segmentation, information retrieval, information extraction, sentence boundary disambiguation, discourse analysis, etc, in our question answering system. We will make use of relevant libraries and modules when available, and implement algorithms as needed.