

## **Initial Plan for 11-411 Final Project**

Team Tune A Fish: Emily Bram, Caitlin Lohman, Ariel Rao, Vijay Viswanathan

### **Parsing the Text**

We will use NLTK to apply part of speech labeling and the Stanford parser for entity relationship modeling. Using an external package to take care of the details of implementation will allow us to focus on tweaking our algorithms on a higher level. We will use these tools to parse text before both distinguishing between the tasks of asking and answering.

Because pronouns are fundamentally ambiguous, we will consider using a probabilistic model for anaphora resolution. Then, we can build an offline database of entity-factoid pairs that we can query for answering.

### **Asking**

Since we have already tagged the text when parsing, we can then identify the candidate subjects of each question in an article and build a collection of question templates. Given a new article, we will consider the attributes of each subject in the text and apply the most probable question template based on all critical words in each sentence.

### **Answering**

Figuring out what type of question (yes/no, location, date, etc) is being asked will be useful for determining which relationships between words we should be considering. Parsing sentences into phrases and then deciding the functionality of the phrase will be useful for answering questions based on types. For example, a prepositional phrase that describes location will be useful for answering a location question while maintaining proper grammar.

If information or relationships in the article are successfully extracted, then this information will be delivered as the answer. Otherwise we will retrieve the most likely sentence, treating the keywords of the question as a vector that we are trying to match, and extract the most salient section of this retrieved sentence as the answer. We will likely use term frequency to rank sentences within a document.

Both the asking and answering modules will share the structured data extracted from parsing text and use wrappers over useful NLP algorithms from NLTK/other libraries. Otherwise the two components' system designs will be independent.

## **Evaluation**

We will be automatically evaluating each question generated for grammar and syntax, to ensure fluency. We will also evaluate answers for surface-level factual accuracy and adherence to the information need of the corresponding question, in addition to grammar and syntax. We will use the quality of candidate answers as a ranking criterion for the questions we generate.

## **Team Coordination**

Our team will be using Git for version control and to share code/data. Because our group is divided among people with a background in programming vs. experience with linguistics, we will be dividing tasks accordingly. One of our team expectations by the first progress report will be to have a full “skeleton” of the functional system, and we will then assign specific coding tasks accordingly. We also have a standing weekly meeting (Thursdays) to delineate tasks and manage our progress.