15-400 Milestone Report III

Sonya Anopa

February 26, 2016

0.1 Milestone Progress

I have finished the is-a link extraction process in KNEXT, which was the stated milestone as modified in the last progress report. The data is pattern-matched to one of the cases described in the paper that accompanied this data and then converted into an appropriate Scone link. These cases are generally equality statements, is-a statements broken up by a classification of the concepts as mass or count nouns, branches of study, and is-not-a links. The concepts created from is-a relationships (and otherwise whenever possible) are created with a superclass of either "countable" or "uncountable" that is identified from the data. Some statements allow for creating a "consists of" relation - e.g. "food consists of nutrients" - which provides additional knowledge over simple is-a links.

I spent a decent amount of time thinking about some cases I wanted to filter out. For example, the original data suggests a physical process is equivalent to a process, which does not make sense. There were also a few such statements involving different English spellings of the same word. For the first case, I decided to identify these more subclass-like statements by simply checking whether one concept name was a substring of the other, at the same time excluding words that came from the same root by stemming. To check for spelling differences, I was initially thinking of some algorithmic way of determining the closeness of the two words (e.g. Levenshtein distance), but then I decided that I could not provide a realiable "closeness" threshold. I found instead a dictionary of common US-UK spellings (because the variations seemed to usually be UK-like spelling differences) and imported it. I use it to decide whether two words are the same word and if so, to add the British spelling as simply another spelling variant of the concept. As it turns out, other statements sometimes include the British spelling as well, so I used this check for all the other cases as well.

I am not completely sure my conversion is correct, so I might tweak it later, but for now I think it is reasonably done. Therefore, I have started looking again at WordNet. I am currently thinking of using a dependency graph to extract some of the relationships between concepts and their definitions; I have identified some binary relations that could be useful, such as adjective modifiers, clausal modifiers of nouns, and nominal modifiers.

0.2 Looking Ahead

I am going to continue trying to extract information from WordNet, hopefully delivering some prototype extracting the types of relations I have described by the next milestone.

0.3 Revisions to Future Milestones

The next milestone predicted "the addition of relations between concepts generated from WordNet information", which is technically still valid, but might not encompass all the relations.

0.4 Resources Needed

I have been looking at the possible resources for dependency graph identification and I think I can use NLTK / the Stanford Dependency Parser, so I think I have all the resources I need so far.