Natural Language Processing – Assessment 3

# Task 1

The verb “use” depends on the root “advised” (indirectly via “you”) – we are therefore looking for a syntactic form involving “use”. By following the dependency from “advised” to “sort” and then to “what” and “water”, we note that we are looking for a type of water. We now want to search for occurrences of the word “use” that have a compound of “water” as their direct object. There are two in the text – “using tap water” and “use distilled water” both have a *dobj* (direct object) dependency from “use” to a compound form of “water”. We’ve now identified the two sentences that contain answer material, and thus we are done.

Looking at the parse for question 12, we see that “extra” is the direct object of “pay”. We’re therefore looking for some kind of syntactic form where “extra” (or a compound form involving it) depends on a form of the verb “pay”. Such a search would include the last sentence of the *Special Diets* section, since this sentence contains a *nsubjpass* dependency involving “paid” and “costs”, and then “costs” appears in a compound form with “extra” – this is not the answer we’re looking for, but the sentence counts as potential answer material.

However, such a search would fail to identify the sentence containing the actual answer (i.e. the last sentence of the first paragraph of the *Accommodation* section), since this sentence uses the term “supplementary charge”. We would need a way of dealing with synonyms – see Task 2.

[248 words]

# Task 2

The parse of question 3 tells us that “drip” is a verb. Looking it up in WordNet, we notice that “drip” belongs in a Synset with “dribble” and “drop”. Each of these verbs has an associated noun, and by inspecting the lexical relations of the word “drop”, we notice the word “droplet”. This indeed appears in the text, and specifically appears in a sentence that answers the desired question (deriving an action from this sentence is somewhat different – Task 3 deals with this).

We could instead use word2vec to find the most similar word in the text, and present this as a candidate sentence. I tokenised Text 1 and checked the similarity of each word against the word “drip” using gensim with the GoogleNews word2vec embeddings. As expected, the most similar was “droplets”, and this gave me a similarity of 0.358.

As we noticed in Task 1, it might be useful to assess semantic similarities between words to help answer Question 12 of Text 2. Words that have similar word2vec vectors to those of “pay” and “extra” might be useful. However, the most similar word to “extra” in the text was “additional”, and this would actually point you to the wrong sentence. “Supplementary” doesn’t even rank in the top 10.

Doing the same with the word “pay”, we obtain the words “costs”, “money”, “charge” and “price” (among others), which seems reasonable. Even then, however, using this approach alone would highlight many possible sentences, leading to noisy results.

[247 words]

# Task 3

We can convert the question into the logical task of finding a usage of the iron that, when performed by a person, would lead to them getting hurt. In the text we notice the statement “*Do not attempt to remove creases from an item of clothing that is being worn*” – negating this produces a possible usage of the iron, but now we need to see if it causes harm. The reasoning chain would take the following form:

* Removing creases from an item implies that we are touching it with the iron.
* The iron is hot, and therefore we are touching the item with a hot surface.
* The sentence tells us that the item in question is an item of clothing; furthermore, the clothes are being worn. Therefore, a hot surface is touching clothes that are worn.
* Clothes worn by a person are in direct touch with the skin, so a hot surface is touching the skin.
* Skin can burn if touched by a hot surface, so the skin risks being burned.
* Burning causes harm.

We’ve therefore found an action involving the iron that causes harm.

The entire task cannot be implemented, however – the largest obstacle is that many of those inference steps require a compositional semantics framework with a knowledge base that encompasses details not included in the text, e.g. burning the skin causes harm to a person. Such details are context-specific, which means that assembling a sufficiently flexible knowledge base is very challenging.

[250 words]

# Task 4

At a high level, we will begin with parsing the question string using a dependency parser and determining which category it belongs to. Judging by the provided question paper, it looks like the types of question might be bounded – there are some wh-questions, some “fill in the blank” questions, some true/false questions (a yes/no question is really just a true/false question), among others. The idea is to hard-code a separate strategy for each type of question we choose to target.

Some ideas are common to all strategies. For example, pruning the search space is crucial to ensure our system isn’t wasting time looking in the wrong places – POS-tagging and word sense disambiguation would help here, as well as using lexical semantics tools (e.g. WordNet) to hone in on relevant sentences by spotting words related to the question.

We’ll also need a compositional semantics framework to perform some inference in the case that the answer is not directly mentioned in the text. The obstacle here is how to assemble the knowledge base (see Task 3), and this could take a large chunk of our time. In any case, lexical semantics tools are crucial to inference steps, since they bridge the gap between logical statements whose meaning is related but the precise words used differ.

We need a heuristic to break ties between potential answers. If answering a multiple-choice question, then we can see which answer fits better with the options. Otherwise, we can try to rank options by looking for comparative/superlative forms of adjectives. Ultimately, the time-investment here will vary – if we struggle to identify the relevant sentences in the first place, we should keep this step simple and instead devote more time to working on our strategies. Post-processing (i.e. re-wording the resulting sentence to sound like an answer) will conform to a similar trade-off.

[304 words]