

## NLP Assignment 3

# Text Understanding

Anik Roy, Christ's (ar899)

December 4, 2019

Total Word Count: 999<sup>1</sup>

## Task 1

A hypothetical system would look at the key verb in the question - here, 'use'. We can find 'use' due to the xcomp dependency from the root. The 'dep' type can then show us what the object of the verb is, here 'water'. The determiner ('what') determines the type of question, and indicates we need to look for a noun and its dependents which are modifiers.

The question *What sort of water are you advised to use?* has the answer *distilled water*, but the question answering system described above would likely return two possible answers - tap water **and** distilled water. In both relevant sentences, 'water' is dependent on a form of the verb 'use' with the direct object (dobj) type. We can take advantage of morphology [] here to detect that 'using' is a form of 'use'. Both 'distilled' and 'tap' are dependents and modifiers of 'water' (amod/compound), as required by the question. To choose a single answer, a simple heuristic is simply choosing the latter. However, we could also look at the other dependents of 'use', i.e. 'last longer', classifying as a positive phrase, making this sentence more likely to contain the correct answer.

Figure 1 shows the dependency parses of the relevant sentences.

The second question is more difficult, since relevant words in the question are not directly present in the text, e.g. 'pay extra' in the question relates to 'supplementary charge' in the text. This requires us to find a way to find similar words and phrases, i.e. calculate semantic similarity. Also, since we have an question that is not a full sentence (by the lack of the 'punct' type), we can compose it with each answer and check for truth.

From the dependency parse of the question, we can see that 'pay' is the relevant verb (via xcomp from root) and 'bathroom' as the object. Since the word bathroom does not appear in the text, we can use a semantic similarity measure to look for the most similar words.

---

<sup>1</sup>Using texcount

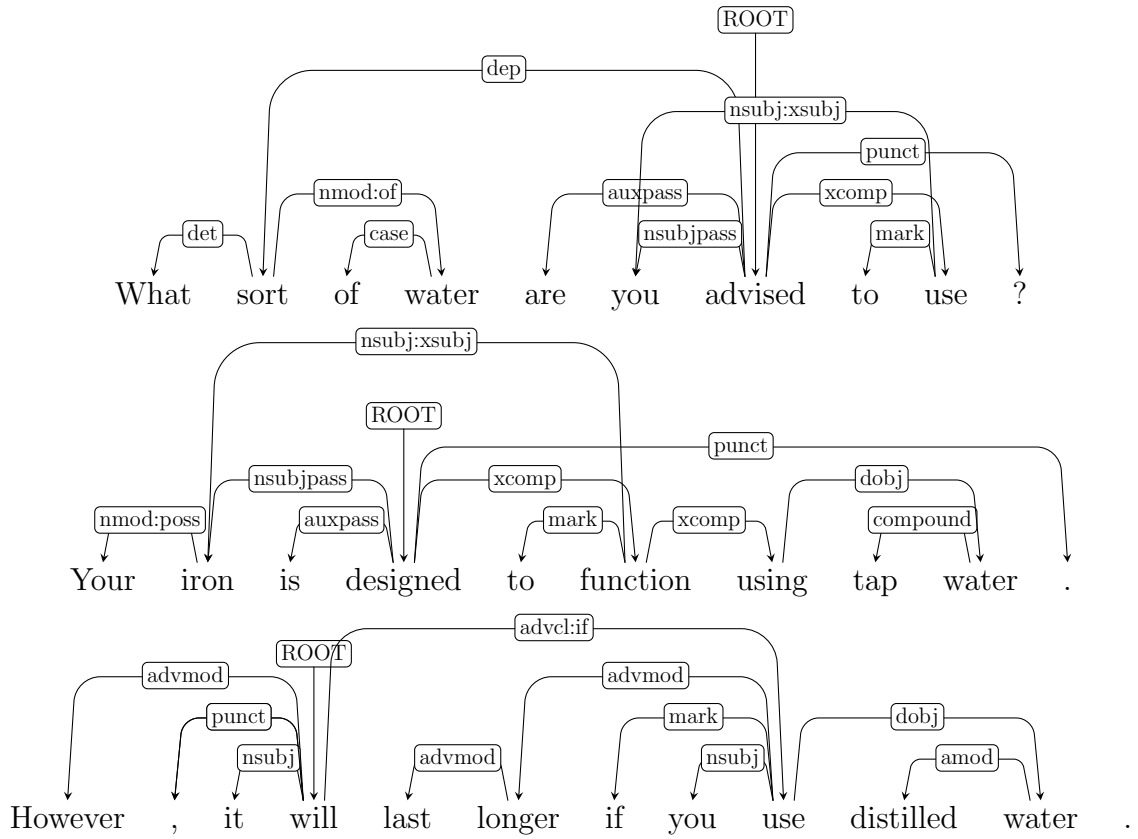


Figure 1: dependency parses relating to question 1 of text 1

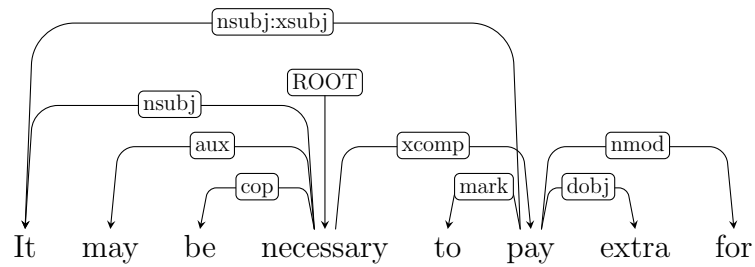


Figure 2: dependency parses relating to question 12 of text 2

## Task 2

Task 1 refers to semantic similarity measures, which need to be used when words occur in questions that don't occur in the text. e.g. 'drip'. I discuss two methods here, wordnet and word2vec, and show that they are able to find that 'droplet' is the most similar word to 'drip', and so the answer to the question lies in the sentence containing the word 'droplet'. Being able to derive an answer to the question from this is more difficult, and requires chains of reasoning (task 3).

Wordnet is a large database of english words grouped into 'synsets', which represent distinct concepts. These synsets are linked by lexical relations. The result is a large network from which related words can be obtained. To use it here, we must first carry out POS tagging [1], since wordnet uses POS information to disambiguate words. The word 'drip' as a verb belongs to 2 synsets, and by traversing derivationally related forms of the words in these synsets, we arrive at droplet in 2 steps, which is the closest word in the text. Shorter chains are one metric for similarity, but other similarity metrics can be employed (Liu et al. 2015) [3].

Word2Vec is another method by which similar words can be found. Word2vec represents words as n-dimensional vectors which are not directly interpretable, using a neural model to learn these 'word embeddings' [2]. This method can also be extended to phrases in a meaningful way [4], enabling us to detect similarity between phrases. Using Word2Vec, we can convert each word found in the text to a word embedding and calculate the cosine similarity - carrying this out on the text shows that the most similar word is 'droplet' (0.390).

Looking at question 12 from text 2, our system must be able to find similarity between the 'pay extra' and 'supplementary charge', as well as 'private facilities' and 'bathroom'. This requires the use of collocations, supported by both word2vec and wordnet - however, neither the pre-trained word2vec embeddings nor wordnet contain these phrases as collocations. In the case of word2vec, more training may overcome this. However, both show similarities between 'pay' and 'charge'.

## Task 3

To answer *Which misuse of the iron could result in a person getting hurt?*, and example of an informal reasoning chain is:

1. Removing creases from implies using the iron
2. Using the iron implies the iron is hot
3. clothes worn by a person are in direct touch with skin
4. Using an iron on these clothes means the iron is in contact with skin
5. Skin can burn if touched by a hot surface
6. So the skin is at risk of being burned

7. Burning is harmful

## References

- [1] Paula Buttery. Nlp lecture 2, 2019.
- [2] Paula Buttery. Nlp lecture 2, 2019.
- [3] X. Liu, Y. Zhou, and R. Zheng. Measuring semantic similarity in wordnet. In *2007 International Conference on Machine Learning and Cybernetics*, volume 6, pages 3431–3435, Aug 2007.
- [4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.