NLP Assignment 3 Report

Investigating an Automated Question Answering System

Zébulon Goriely, Queens', zg258

Thursday 5th December, 2019

Introduction

For all three tasks, I investigated the NLP problem of understanding text. The specific scenario I undertook is the simulation of a second language learner's performance on the IELTS test[1]. These tests involve reading provided text and answering non-trivial questions, specifically designed to require inference.

Example questions were provided in pdf format, plaintext format and parsed format using the Stanford Parser [Chen and Manning, 2014] in the framework of a course in NLP.

Task 1

Word Count: 332[2]

The first provided text gives an instruction manual for a Moulex Iron. The first question associated with this text is *"What sort of water are you advised to use?"*. When answering, I implicitly:

- Realise that the "what kind of" question implies that the answer is a subtype of water
- Realise that from "advised to use", the answer is in the context of "use"
- Narrow the answer down to "distilled water"

I can translate this intuition into a hypothetical automated system. To operate on the text and questions, I use dependency parsing [Buttery and Copestake, 2019b]. Dependency parsing produces structures that are intuitively closer to the meaning of the text than regular parse trees. It also has the advantage of being neutral to word order.

The hypothetical system restricts answers based on dependencies and begins by analysing the parse and part-of-speech tagging of the question, as seen in Figure 1. I then imagine a sub-system that uses this information to derive the question type; the `WP` tag signifying a "wh-" question. The `nmod:of` and the `nsubj:xsubj` dependencies in the parse informing the system that answer is a subtype of "water" used in the context of "use".

The system then compares each occurrence of "water" with the parse structures, searching for the dependencies `compound(water, xx)` or `amod(water, xx)` and `dobj(use, water)`. This may involve stemming [Buttery and Copestake, 2019a] in order to accept other forms of verbs and nouns. The answer to the question is then `xx water`.

Figure 2 and Figure 3 show the two solutions that this system finds, "tap water" and "distilled water". I can use a heuristic to select the second choice, intuitively because detail is likely given later in informative text.

[1] https://www.ielts.org/

[2] *texcount docs/assignment3/report.tex*

This imaginary system fails, however, when applied to Question 12 of Text 2. Neither the phrasing of the question nor the actual answer ("bathroom") appear in the text, so a search for `nmod:for(extra, bathroom)` fails. The inference must be made on the word "facilities", indicating the need for a more complex system.

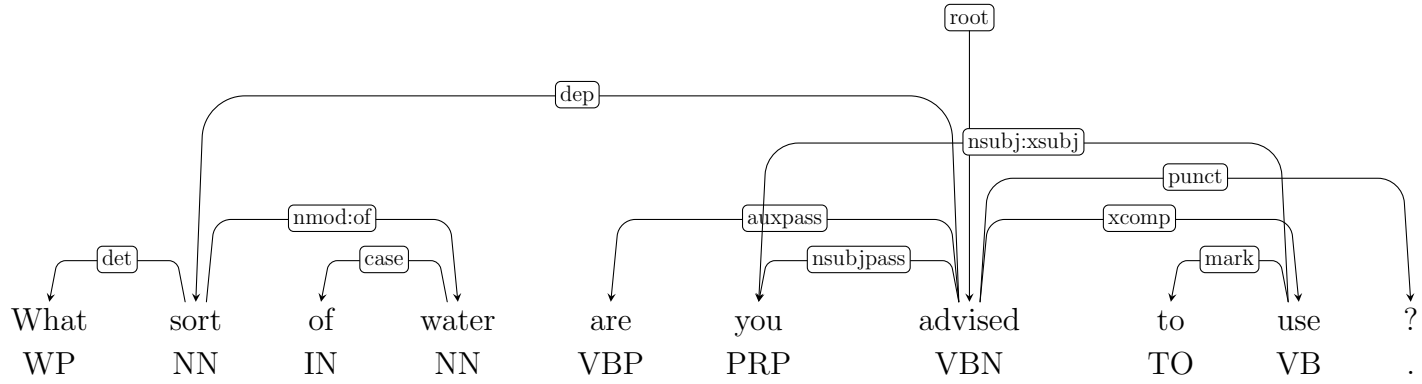

Figure 1: Stanford parse and part-of-speech tagging of Question 1 from Text 1

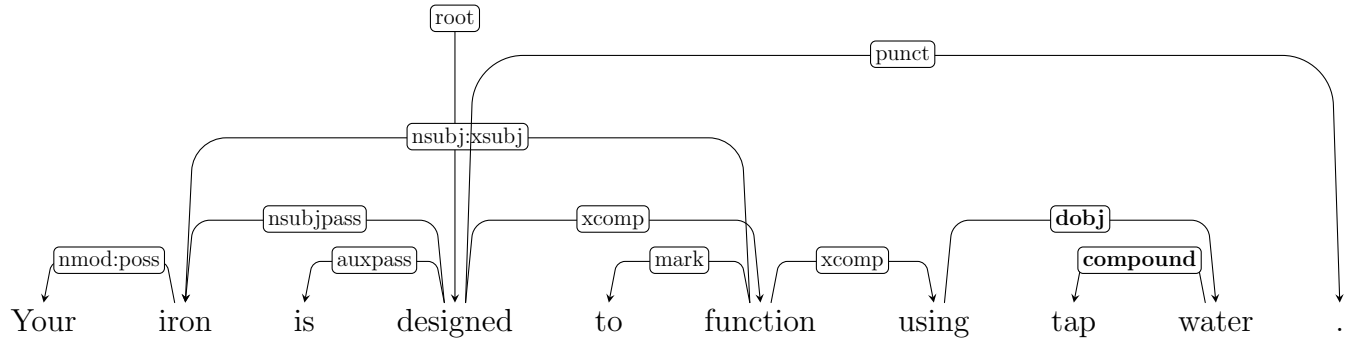

Figure 2: Stanford parse of Sentence 2 from Text 1 with search-tags in boldface

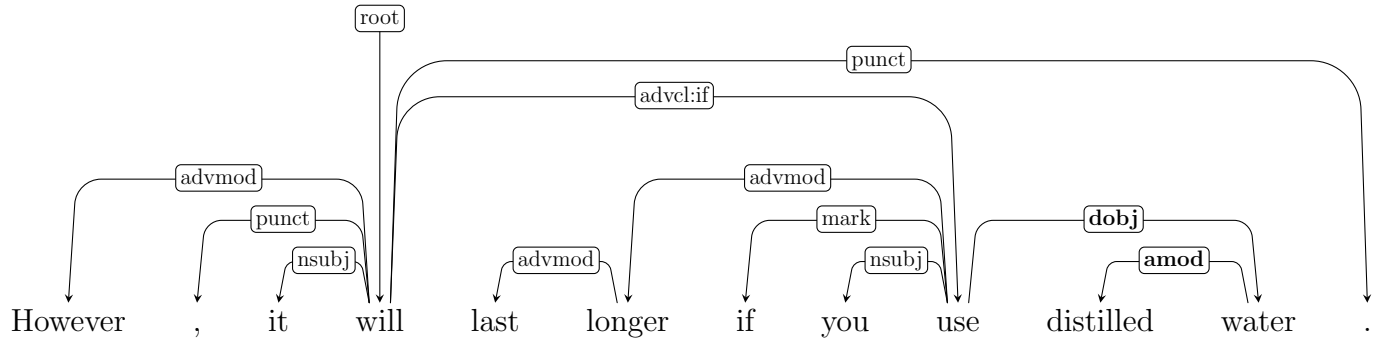

Figure 3: Stanford parse of Sentence 3 from Text 1 with search-tags in boldface

Task 2

Word Count: 332

I imagine the system attempting to answer the question *"What should you do if your iron starts to drip water?"* for the first text.

As before, the system uses the dependency parse and part-of-speech tagging of the question (see Figure 4) to restrict the possible answers to conform to the dependencies `dobj(drip ,water)` and `advcl:if(xx, drip)`. The answer will relate to the verb `xx`.

Unfortunately, such a verb does not appear in the text. The system can be improved by using lexical semantics [Teufel and Copestake, 2019b] to rank potential answers by lexical similarity.

I explore three different metrics for lexical similarity. Firstly, using pre-trained `word2vec` embeddings provided by Google [Mikolov et al., 2013], I find that "droplets" is the second-most similar word in the text to "drip", as seen in Table 1. Similarly, exploring the synset of "drip" using *WordNet* [Miller, 1995] gives "drop" which is found contained in "droplet". Note that I use the noun form of "drip", since words that are both nouns and verbs can always be modified by changing the tense or voice. Finally, using *ConceptNet* [Speer et al., 2016] relations (see Figure 5) gives another means of deriving that "droplet" is one of the closet words in the document to "drip".

Similarly for Question 2, the phrase *"quantity of steam"* does not appear in the text. Performing a `word2vec` search (see Table 2) shows that "amount" is the most similar word to "quantity", telling the system that the answer lies within the sentence containing "amount of steam". This is also supported by *WordNet*, the synset for "quantity" containing "amount" (see Figure 6).

Combining the lexical ranking of "drip" with our dependency parsing lets the system derive that the sentence *"If your iron produces droplets of water instead of giving off steam, your temperature control is set too low"* relates to the answer of Question 3. The `advcl:if(set-20, produces-6)` dependency gives "set" as the verb required. Further inference is required, however, to derive that "temperature control is set too low" implies that the answer is "increase temperature" which I discuss in task 3.

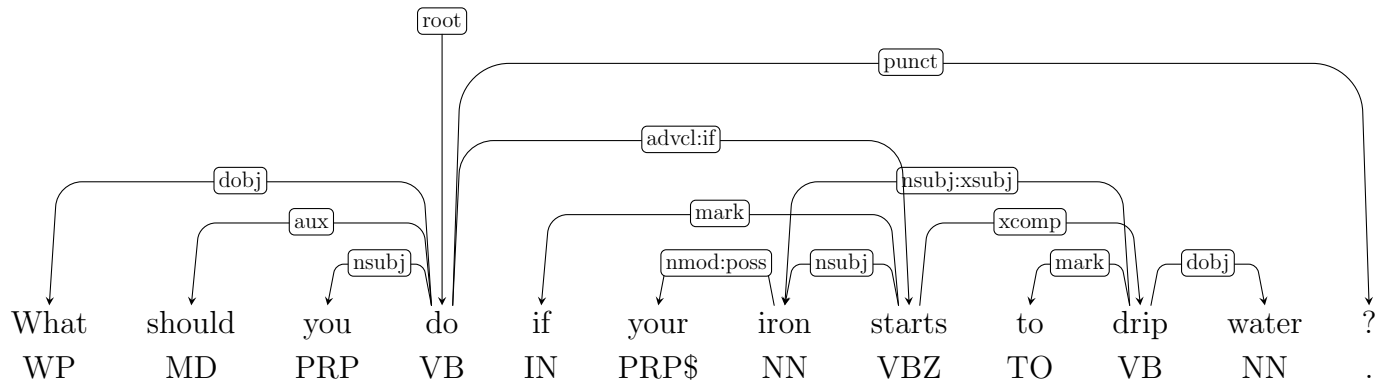

Figure 4: Stanford parse and part-of-speech tagging of Question 3 from Text 1

Word	Similarity
“drip”	1.0
“spray”	0.44
“droplets”	0.36
“distilled”	0.33
“water”	0.32

Table 1: The 5 closest words in Text 1 to “drip” according to `word2vec`

Word	Similarity
“quantity”	1.0
“amount”	0.44
“intensity”	0.28
“therefore”	0.28
“produced”	0.26

Table 2: The 5 closest words in Text 1 to “quantity” according to `word2vec`

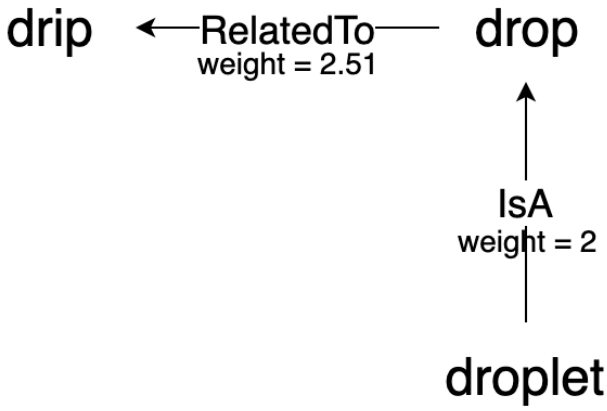

Figure 5: ConceptNet derivation of “droplet” from “drip”

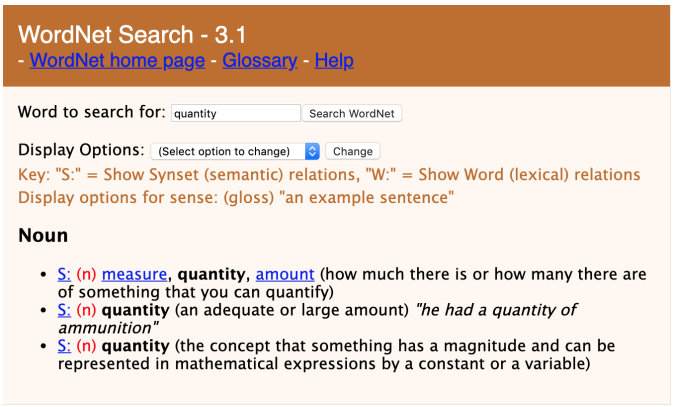

Figure 6: WordNet search for “quantity”

Task 3

Word Count: 326

Sometimes, answering questions requires logical reasoning. Question 5 of Text 1 requires the following chain of reasoning:

- Irons are used to remove creases
- Do not attempt to remove creases from an item of clothing that is being worn
- Clothes worn by a person are in direct touch with skin
- Skin can burn if touched by a hot surface
- Irons are hot surfaces
- Burns hurt
- Ironing clothes being worn can results in a person getting hurt

Automating a chain requires inference on a knowledge base [Teufel and Copestake, 2019a]. For this task and other NLP tasks, common sense reasoning a major obstacle to achieve similar performance to humans. To explore if this automation is possible, I explore how ML is used to solve similar tasks.

Two popular datasets for the common sense reading comprehension task are the *Story Cloze Test* [Mostafazadeh et al., 2017] and *SemEval-2018 Task 11* [Ostermann et al., 2018]. Like Question 5,

some of the knowledge required to answer the questions may not be found in the documents, requiring systems to be equipped with some common sense knowledge database. These datasets may be effective for training a system to answer IELTS questions.

Many publicly available knowledge sources already exist, such as *ConceptNet* [Speer et al., 2016], *WebChild* [Tandon et al., 2014] and *DeScript* [Wanzare et al., 2016].

Wang et al. [2018] have shown using *ConceptNet* as a knowledge base can achieve accuracies of 83.95% on *SemEval-2018 Task 11*, but their system only uses embeddings based on *ConceptNet* as additional input features. They propose that methods based on event calculus [Mueller, 2014] are more rigorous mathematically and more closely resemble humans processing.

Performance on these tests has rapidly improved, machine learning models such as in [Xia et al., 2019] achieving accuracies of 88.23% and 87.4% on *SemEval-2018 Task 11* and *Story Cloze Test* respectively. However, these models make it much more difficult to justify the intermediate steps.

This rapid improvement suggests that achieving an automated system that matches human performance on the IELTS test by 2021 seems possible. Whether or not such systems will be able to justify their answers, as required by the specification, remains to be seen.

References

Buttery, P. and Copestake, A. (2019a). Natural language processing: Part ii overview of natural language processing (l90): Acs, lecture 2: Finite-state techniques.

Buttery, P. and Copestake, A. (2019b). Natural language processing: Part ii overview of natural language processing (l90): Acs, lecture 5: Dependencies.

Chen, D. and Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mostafazadeh, N., Roth, M., Louis, A., Chambers, N., and Allen, J. (2017). Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51.

Mueller, E. T. (2014). *Commonsense reasoning: an event calculus based approach*. Morgan Kaufmann.

Ostermann, S., Roth, M., Modi, A., Thater, S., and Pinkal, M. (2018). Semeval-2018 task 11: Machine comprehension using commonsense knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 747–757.

Speer, R., Chin, J., and Havasi, C. (2016). Conceptnet 5.5: An open multilingual graph of general knowledge.

Tandon, N., De Melo, G., Suchanek, F., and Weikum, G. (2014). Webchild: Harvesting and organizing commonsense knowledge from the web. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 523–532. ACM.

Teufel, S. and Copestake, A. (2019a). Natural language processing: Part ii overview of natural language processing (l90): Acs, lecture 6: Compositional semantics.

Teufel, S. and Copestake, A. (2019b). Natural language processing: Part ii overview of natural language processing (l90): Acs, lecture 7: Lexical semantics.

Wang, L., Sun, M., Zhao, W., Shen, K., and Liu, J. (2018). Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension. *arXiv preprint arXiv:1803.00191*.

Wanzare, L. D., Zarcone, A., Thater, S., and Pinkal, M. (2016). Descript: a crowdsourced corpus for the acquisition of high-quality script knowledge. In *The International Conference on Language Resources and Evaluation*.

Xia, J., Wu, C., and Yan, M. (2019). Incorporating relation knowledge into commonsense reading comprehension with multi-task learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2393–2396. ACM.