Assignment 8: Evaluating Parsers

1 Evaluating a parser

You have been hired by a AwesomeSearchEngine.com to make recommendations concerning language technology. They have been approached by Whizdee, Inc., a startup company, and in a sales presentation Whizdee's sales representative says the following: "We're doing very exciting work on parsing, and our results are very impressive. In one experiment, we trained one of our parsers on 80% of the Penn Treebank and tested on the other 20%, and its labeled recall on constituent boundaries was 98.2%. We also did an experiment, training on the same data, where the labeled precision for constituent boundaries was 97.6%. With numbers like that, how could you lose?"

(a) Consider the true parse T, and the system parse P, below:

```
[T] (S (NP the prof) (VP (VPRT looked up) (NP the grade)) (ADVP today))
[P] (S (NP the prof) (VP looked (PP up (NP the grade)) (ADVP today)))
```

What are the values for labeled precision and labeled recall? Note that I've omitted all part of speech labels because they're not used in constituent recall/precision calculations. Every uppercase symbol in the parse is a constituent label.

(b) Should people be impressed by Whizdee's numbers? Explain why or why not.

(c) Whizdee's decided you're so smart that they, too, want to pay you as a consultant. They've got a question-answering search engine that uses a parser, and a contract from the New York Times. They paid minimum wage to impoverished linguistics grad students in New York City to create parse trees for 20,000 New York Times articles written between March 15, 2007 and March 15, 2017. Whizdee plans to make their system available to the public starting April 1, but the New York Times insists on a formal parser evaluation first. One of Whizdee's scientists says that they should evaluate their parser by doing 10-fold cross validation. Another of their scientists says that they should evaluate it by training on the data up to March 15, 2016 and testing on the rest. Explain what the two competing evaluation approaches are and briefly discuss the advantages and disadvantages of each approach.

2 Evaluating a research problem

Suppose the CEO of AwesomeSearchEngine.com approaches you and says the following: "We are considering having two of our full-time PhD-level researchers focus on the problem of parsing Bantu languages, using Whizdee's annotation-and-training approach, and we would like you to help us evaluate whether or not this is a good research problem for us." Adopting the perspective of Cohen and Howe's framework for evaluation guiding AI research (AI Magazine, Volume 9 Number 4, 1988, Figure 1), how would you help them answer their question?

Your response to this question should address each of the six criteria that Cohen and Howe lay out. For some of these, you might have plausible answers to suggest. For others, the best response might be to identify questions that need to be answered first, before you can make an informed assessment of the project.

3 Looking at real-world performance

Pick an off-the-shelf parser, e.g. NLTK, spaCy, the Stanford Parser. For this assignment, you can use online parser demonstrations (e.g. Stanford at http://nlp.stanford.edu:8080/parser/, sPacy at https://demos.explosion.ai/displacy/), or you can install and run the relevant code yourself.

Pick at least ten real-world sentences at random from at least two diverse sources, e.g. Google News, Twitter, the UMD Web site, the U.S. Congressional Record, ... Run the parser you chose on those ten sentences, and briefly assess, manually, the quality of the information the parser is giving you. Is it giving you useful information? What kinds of mistakes is it making, and under what circumstances would those mistakes matter or not matter?

Extra credit (up to 5%): Do the above with a second parser, and compare/contrast the outputs of the two parsers.