Implementation

The skeleton code on the course web page is the start of a program for parsing sequences of tokens according to a context-free grammar. The code includes a hopelessly inefficient parsing algorithm that tries all possible parses until finding one that matches. Implement a dynamic programming approach for parsing, and compare its performance to the enumeration approach that we provide. Those in 758 need to return all possible parses. Those in 858 must implement probabilistic parsing and merely return the most probable parse. In this context, the DP approach is called ‘the Cocke-Kasami-Younger algorithm’ or ‘CKY.’

The skeleton code will parse the input string into an array of tokens. Your output must include the parse in bracketed notation, so you will need to keep track of which non-terminals and spans are used in the parses you construct.

The grammar (in Chomsky Normal Form) is formatted as:

S NP VP 1
VP V .386
VP V NP .393
VP V NP PP .22
NP NP PP .24
NP N N .09
NP N .14
NP ART N .55
PP P NP 1
P about 1
NP fish 1
N fish 1
V fish 1
VP fish 1
ART the 1

The start symbol for the grammar is required to be S. All nonterminals are required to be in capital letters. All terminals are required to be in lower case. The number following each production is the probability associated with that production.

In the above grammar, the word "fish" can be used as either a noun or a verb, and as a result, there are quite a few valid parses for the sentence fish fish fish. Some selected examples include:

S(NP(fish) VP(V(fish) NP(fish))) 0.007703
S(NP(N(fish) N(fish)) VP(fish)) 0.034740

The number is the probability associated with that particular parse.

Undergraduates are expected to output all valid parses of the sentence, and do not need to output any probability. Graduate students are to output only a single parse, the most probable one, along with the associated probability.

Testing

parser The skeleton can be run as ./parser grammar.cnf <alg> where <alg> may be either rdp (recursive descent parser) or cyk, which is the name of the dynamic programming parsing algorithm

harness.py The harness for this assignment is a python script. Usage:

The arguments are:

- `-h, --help` show usage and exit
- `-binary BIN` binary to use
- `-ref REF` reference solution to use
- `-sentence SEN` sentence to try and parse
- `-grammar GRAMMAR` grammar to use
- `-grad` whether or not to use graduate student mode

The harness checks the output from the student solution and the reference solution, and outputs any errors it finds. When it is done, it writes the word "Done". If this is the only output it produces, it found no problems, otherwise it will output information about any discrepancies as well.

**generating grammars** There is a utility java program that can be used to convert context free grammars to CNF. It comes as a runnable jar, as well as a collection of java files.

It can be run as `java -jar cfg_convert.jar data/english.cfg`.

This will take the grammar in `data/english.cfg` and convert it to a grammar in Chomsky Normal Form. In the process of adding grammars, it will often be required to add its own symbols. It uses the prefix ADDED followed by a letter, so don’t use nonterminals like that in the original grammar!

**Written Problems**

1. Briefly list any parts of your program which are not fully working. Include transcripts or plots showing the successes or failures. Is there anything else that we should know when evaluating your implementation work?

2. What is the time complexity of your program? How does this compare to the enumeration approach? Is this reflected in the plots?

3. (Those in 858 only) Problem 15–6 in CLRS. If you get totally stumped, one design technique is to start with exhaustive enumeration, then memoize, then convert to DP.

4. Exercise 16.2–1 in CLRS.

5. Exercise 16.2–4 in CLRS.

6. What suggestions do you have for improving this assignment in the future?

**Submission**

Electronically submit your source code using the script on agate (eg, `~cs758/scripts/sub758 8 your-asn8-dir`). Also hand in a listing of your source code (2 pages per page, as with `a2ps -2`), along with your written work, to the TA in class.

**Evaluation**

In addition to correctness, your work will be evaluated on clarity and efficiency.

Tentative breakdown:

- 7 parsing
- 3 written problems