

6.863J/9.611J Reading and Response 3

Genesis and Implementation: stochastic context-free grammars.

Goal: how to write for Wikipedia; how to research and summarize intellectual history

Handed out: Wednesday, March 19th

Due: Wednesday, April 2, for in-class discussion. (*Note* unusual day – Wednesday, not Monday!)

Readings. Available via links on the course home page for the Readings on 3/12, as well as under the “Announcements” at the head of the home page and on the “Assignments” page:

1. Carl deMarcken, “Lexical heads, phrase structure, and the induction of grammar,” *Proceedings of the 1995 Meeting of the Association for Computational Linguistics, Sigdat*.
2. Michael Collins, “Head-driven statistical models for natural language processing,” *J. Association for Computational Linguistics*, 2003.

Both papers may be challenging to read. To make your job simpler, for this assignment you only have to “skim read” them to get the key ideas. In particular, the first paper talks about the “EM” algorithm for estimating the probabilities associated with a stochastic CFG, a general local search method for maximizing probability estimates. If you haven’t run across EM before (we shall cover it a bit later), then (surprise!) the Wikipedia entry is not a bad summary:

http://en.wikipedia.org/wiki/Em_algorithm

(Historical personal note: The method was invented by Art Dempster at Harvard, and then published as a joint paper a few years after. I recall this quite clearly since I was sitting in Art Dempster’s statistics class the day he walked in and announced that he had figured out this method, which he then proceeded to outline on the board...) The second paper is especially long (48 pages). We do not want you to have to read all of it. Skim the following parts of this paper (you can read the rest if you want): section 1 (introduction); sections 2.1 and 2.2 (about probabilistic context-free grammars); just the first two paragraphs each in sections 3.1 (‘Model 1’) 3.2 (‘Model 2’), and 3.3 (‘Model 3’); section 6 (Results); sections 7.1 through 7.3; and section 9 (Conclusions).

Your Task. You have been miraculously transported to the year 2050 and are working on the Hitchhiker’s Guide to the Galaxy. (Thankfully, by then Megadodo Publications has bought out both Google and Wikipedia. It was no contest – they threatened to destroy Earth, or something like that. It’s hard to remember). As part of your assignment, you are reviewing the entry that recites the history of stochastic context-free grammars and their use in natural language. Today you’re trying to get straight the intellectual relationship between the two readings handed to you above.

Your job is to write a Wikipedia entry for just these two articles, in the style of an encyclopedia. For reference, you look at a Wikipedia entry, e.g., the entry on ‘stochastic context-free grammars’, http://en.wikipedia.org/wiki/Stochastic_context-free_grammar

Feel free to look at other short Wikipedia entries to get a feel for the style. (You download the web page and archive it to hack it in creating your submission, if you want.)

Since there are a lot of other articles to cover in the Galaxy, you must limit your written response to a two pages and follow the usual style guidelines. Please email your pdf, plaintext, or, in the best case, a Wikipedia-like html-formatted response, to me, berwick@csail.mit.edu, preferably the day before class; bring a hardcopy to class as usual.

In writing your entry, please answer the following questions specifically, working them into your text. With respect to the deMarcken paper:

- (1) deMarcken argues that there is both a central representational problem with training stochastic context-free grammars and a computational problem. In your own words, and as succinctly as possible, describe what these two problems are.
- (2) Provide a simple linguistic example, similar to the one that deMarcken exhibits, illustrating the representational problem.
- (3) Finally, summarize what deMarcken calls for as a solution to these two problems.

With respect to the Collins paper:

- (1) What parts of deMarcken's representational proposals does it adopt? Be specific.
- (2) Which proposals are novel to the paper itself? Please describe these via a short list. (Here we mean specific ideas and methods that are introduced here for the first time, as opposed to being algorithmic methods referenced in the paper itself – for example, the paper notes that the method of using 'histories' as a conditioning probabilistic context was first used by Black at IBM. How was this modified?)
- (3) How well does this method succeed or fail as an implementation of deMarcken's proposals? What remains to be done, if anything? What improvements were made? How is it different from deMarcken's proposal? In particular, consider the example that you suggested for the deMarcken paper item (2), and see if the Collins approach solves it. For this it may be useful to look at sections 7.1–7.3 of the Collins paper. Please illustrate your conclusions with concrete examples.