

Final work in the *Topics in Natural Language Processing* course

Vassilii Khachaturov

March 12, 2006

Abstract

Probabilistic Context Free Grammars and probabilistic parsing are briefly introduced, following the book of Manning and Schütze ([MS], pp.381–455). Two probabilistic parsers, `dbparser` and `MINIPAR`, are then evaluated on a subset of the WSJ corpus from the Penn Treebank according to the tree accuracy criterion. The sentences selected were the most difficult ones in the corpus, according to three empirically chosen criteria — parse tree width, height, and the top-level verb ambiguity. The two parsers perform similarly on the first two groups, while on the third one `dbparser` is significantly better. Further possible comparison directions are then outlined.

1 Introduction

Context Free Grammar (CFG) is a useful model for capturing language structure, in both parsing and generation applications. However, some rewritings of a CFG for a particular language are more likely than others. Thus, when a parser based on the CFG has to robustly choose between several rewritings, its internal model needs to be enriched with some additional context-related data, such as a probabilistic model following the language specifics. The original CFG is then called a *backbone CFG*.

Probabilistic CFG (PCFG) (see [MS]) is a CFG in which for every non-terminal N^i and all its rewriting rules of the form $N^i \mapsto \zeta^j$, there is a probability attached to each such rewriting

$$\forall i \quad \sum_j \Pr(N^i \mapsto \zeta^j) = 1 \quad .$$

PCFGs are robust with respect to coping with real text corpora, that have some small share of grammatical mistakes and other errors. In such cases PCFGs assign implausible sentences a low probability. While PCFGs do give some ideas on the plausibility of different parses, in practice, they give a worse language model for English than an n -gram model ($n > 1$) — an n -gram model takes into account the local lexical context, while a PCFG does not. PCFGs also suffer

from certain biases rooted in a specific backbone CFG structure, disagreeing with real corpus data statistical features.

On the other hand, some errors encountered with the n -gram model are clearly rooted in the lack of syntactic structural knowledge of that model. Thus, while PCFGs are not good models by themselves, one is looking for better approaches that combine their strengths with those of an n -gram model.

Part of the toolset used in finding the probability of a given sentence under a grammar G are the *inside* and *outside probabilities* of a sequence of words w_p, w_{p+1}, \dots, w_q . The inside probability $\beta_j(p, q)$ is defined as the total probability of generating the sub-string $w_p \cdots w_q$ given that one begins the derivation with the non-terminal N^j . The outside probability $\alpha_j(p, q)$ is the total probability of beginning with the start symbol N^1 and generating the non-terminal N_{pq}^j (the one that spans the positions p through q in the generated string) and all the words outside these positions. The *inside algorithm* and the *outside algorithm* utilise the dynamic programming techniques to compute the inside and outside probabilities and thus find a probability of a given full sentence under the given grammar.

The *Inside-Outside Algorithm* (IOA), based on the inside and outside probabilities computation, can be used to fine-tune the probability parameters in a PCFG. The algorithm allows to train the PCFG given a representative corpus as an input. The output is a PCFG under which the corpus is likely to occur. The IOA is iterative, improving the likelihood of the observed corpus with each iteration by adjusting the PCFG parameters. Each iteration of training for each sentence takes $O(m^3n^3)$, where m is the length of the sentence, and n is the number of non-terminals in the grammar. The search performed by the algorithm is prone to being stuck in local minima, and the resulting grammar usually is far from the familiar structures occurring in grammars hand-written by linguists.

To train a probabilistic parsing engine on real data, collections of pre-parsed data, a.k.a. *treebanks*, have been made available.

Probabilistic models for estimating a likelihood of a sentence are either *parsing models* or *language models*. The former are based on probability distributions over all the parse trees yielding a given sentence under a given grammar. The latter, rather than using probabilities conditioned on particular sentences, are based on a distribution assigning probabilities to all parse trees (under the given grammar) possible in the entire language. Language models appear to provide a better foundation for modelling than parsing models.

In PCFG-based probabilistic models, due to its context independence, the probabilities of particular parse trees do not take into account the *priming* effect (that the previous conversational context influences a real life person's parsing). Another weakness stemming from the independence assumptions of PCFG-like models is their lack of *lexicalization*. E.g., the chance of a verb phrase expanding as a verb followed by two noun phrases is independent on the actual verb involved under such a model, whereas in reality a ditransitive verb is much more like to have such an expansion. A common way to lexicalize a CFG is to mark each phrasal node symbol by the head word, and define the

probabilities based on such marked symbols. Another weakness of the PCFGs is that probabilities actually dependent on structural context — the likeliness of a given expansion of a particular non-terminal will depend on the context of that non-terminal in the containing tree. These weaknesses suggested building probabilistic parsers better than the ones based on a PCFG, by taking into account lexical and structural context.

While the PCFG model corresponds to a probabilistic version of top-down parsing (at each stage child nodes are predicted conditioning only on the knowledge of the parent node), other parsing algorithms suggest different models of probabilistic conditioning, with each choice point in the algorithm assigned some probability, conditioned on the current state or some portion of the current state.

This includes Probabilistic Left-Corner Grammars (PLCGs), based on the left-corner parsing algorithm (a mix of top-down and bottom-up parsing approach); PLCGs have been shown to outperform basic PCFGs significantly. Other works create probabilistic models based on the choice points of bottom-up shift-reduce parsers.

Another approach to describe the phrase structure is by using dependency grammars. There are isomorphisms between various types of dependency grammars and the corresponding types of CFGs. Dependency grammars are better suited for disambiguation decisions made directly in terms of word dependencies in a phrase, with no need to make such decisions at phrase structural level well away from the words. This approach is thus better suited for chunked parsing, where it's enough to understand dependencies between several syntagms. Also, dependency grammars provide richer toolset to cope with unseen tree patterns (especially rare flat trees), where there is not enough learnt knowledge about exactly such tree structure. While PCFG-based model would have to back off to some default “unseen tree” probability for an unseen local tree, a dependency grammar allows to decompose such tree into several simpler known dependency structures, and, with further cross-conditioning knowledge of such models learnt from the corpus, to achieve robust results in such cases.

For a particular parser, one is interested in having a way to measure the parsing performance. In addition to general algorithmic performance measures, such as the running time and memory usage, a way to measure the accuracy is needed. With the existing treebanks providing pre-tagged parse trees of the sentences therein, for a given sentence one can compare the “gold standard” parse tree from the treebank with the output of the parser under the evaluation. Several criteria for such measurements exist. The strictest criterion, known as the *tree accuracy* or the *exact match* criterion, is to award the parser a binary (0/1) score for each sentence based on whether the trees match exactly. *PARSEVAL measures* are able to award sort of a partial credit for an inexact match, by evaluating the component pieces of a parse. Under the *PARSEVAL* measures, the parser is scored on *precision* — how many brackets in the parse match those in the correct tree, *recall* — how many brackets of the correct tree appear in the parse, and *crossing brackets* — how many constituents in one tree cross over the constituent boundaries in the other tree.

Dekang Lin argues [DL] that dependency-based evaluation brings more meaningful results for scoring of partially correct parses. However, the difficulty of doing so is that information on the correct dependencies in a sentence may not be readily available in a treebank.

2 Parsers compared and the data used

In this work we try to compare two parsers — `dbparser` and `MINIPAR` — using the Penn Treebank WSJ corpus. These parsers are not exactly in the same league to be compared with regular precision and recall measurements.

First, the parsers assume two different kinds of input. `MINIPAR` only operates on untagged phrases. On the other hand, `dbparser` is tuned to operate on pre-tagged inputs. It’s possible to use `dbparser` on regular untagged phrases, requesting the parser to do the part of speech tagging internally, but the settings files that direct the internal decision-making provided with parser had been fine-tuned to the case when an external POS tagger is used as a pre-processor. We did our comparison using the internal POS tagging in `dbparser`.

Second, the parsers are producing different kinds of output. While `dbparser` yields a full syntactic parse tree in the Penn Treebank format (which can be validated against the pre-tagged parses available in the Penn Treebank), `MINIPAR` only gives a dependency-based parse for a given sentence. Thus, the bracket structure is different in the two parsers’ output. Inferring one from the other in a non-ambiguous way is a task of a magnitude beyond the scope of this project; therefore, rather than automatically compare these parsers on a big bulk of data, and gaining statistics on their average relative performance, we opted for a different approach. Three different empirically-chosen criteria were used to select 15 of the most difficult sentences per each of the criteria (using a set of Perl scripts written specifically for this purpose) — tree height, tree width (number of the top-level constituents), and the top-level verb ambiguity. The parsers were run on these sentences and the feasibility of their output evaluated manually, according to the tree accuracy criterion.

Finally, `MINIPAR` is not re-trainable, and was suited to cope with a broader thematic corpus than the Wall Street Journal articles in the Penn Treebank. `dbparser`, on the other hand, is re-trainable, and for the purpose of our evaluation was trained on the very corpus subset as suggested in [UG] — WSJ volumes 02 through 21. The remaining subset of the WSJ corpus, i.e., volumes 00, 01, and 22 through 24, were then used to select the seemingly difficult sentences as described above.

3 Results

On these 45 sentences, `dbparser` took close to 1 minute per sentence on the “highest” difficult group alone. On the other hand, `MINIPAR` spent less than 3 seconds overall for all the 3 groups of difficult sentences combined, when run on

the same Linux workstation.

The tree accuracy results for each of the three difficult sentence group are detailed in the tables 1, 2, and 3. Each of the entries therein consist of a phrase score (the one used for selecting the sentence in the respective difficulty group), the phrase proper, and the accuracy key. The accuracy key displays **D** for a full match by `dbparser`, and **d** for a full structural match with the only mistakes being the part of speech tagging. (Note that the POS tagging errors were all consistent with the known difficult cases described in the Penn Treebank tagging guide.) When a full structural match of the dependency tree was observed for the `MINIPAR`, **M** was written down in the table (note that no POS tagging verification has been done). Additionally, **?** was used to mark the sentence on which the parser failure was coinciding with the attachment of the mis-tagged quoted material, and **†** was used twice to mark a “worse than a failure” condition, when the `dbparser` failed by crashing, no matter how high the heap size was set. Absence of a mark for a parser indicates that it fails the exact match criterion, yet produces some kind of output tree.

One can see that for the “widest” sentences, the parsers behave more or less the same, parsing approximately half the group; with `dbparser` behaving better on FRAG-type phrases which read more like table annotations, and `MINIPAR` once managing to understand a complex conjunction which `dbparser` couldn’t parse properly. It is also in this group that `dbparser` had crashed twice (however, the two sentences were unparsable by `MINIPAR` as well).

The “highest” sentences proved the most difficult for both parsers. Here each parser managed a single (different) sentence out of the whole group, and `dbparser` had a POS tagging difficulty as well.

Finally, on the sentences with the most ambiguous top-level verbs, `dbparser` is clearly outperforming `MINIPAR`, at thirteen matches (four of them with POS-tagging difficulty) against just four; only one of the latter four sentences is not among the former thirteen ones.

4 Open issues

The introduction should probably be updated to the cutting-edge probabilistic parsing techniques. The parsers compared could be compared on other kinds of difficulty groups, such as the “and” conjunctions, in furtherance of the recent research by Yoav Goldberg [YG]. An attempt for automatic harmonisation of the Penn Treebank standard parse trees and the `MINIPAR` dependency trees could be made, and then the parsers could be easily cross-evaluated on a much larger corpus. Finally, the `SUSANNE` corpus (in a sense, `MINIPAR`’s home ground) could be used as well as a source for the sentences to cross-evaluate the two parsers.

A Tree accuracy detailed results

18	Victor Stanley Fishman , Longwood , Fla. , fined \$ 25,000 ; William Harold Floyd , Houston , \$ 100,000 ; Michael Anthony Houston , Bronx , N.Y. , \$ 15,000 ; Amin Jalaal-walikraam , Glenham , N.Y. , \$ 60,000 ; Richard F. Knapp , London , \$ 10,000 and 30-day suspension ; Deborah Renee Martin , St. Louis , \$ 15,000 ; Joseph Francis Muscolina Jr. , Palisades Park , N.J. , \$ 15,000 ; Robert C. Najarian , Brooklyn Park , Minn. , \$ 15,000 ; Edward Robert Norwick , Nesconset , N.Y. , \$ 30,000 .	d
18	Charles D. Phipps Sr. , Hermitage , Pa. , fined \$ 10,000 ; David Scott Rankin , Lake St. Louis , Mo. , \$ 15,000 ; Leigh A. Sanderoff , Gaithersburg , Md. , fined \$ 45,000 , ordered to disgorge \$ 12,252 ; Sandra Ann Smith , Ridgefield , N.J. , \$ 15,000 ; James G. Spence , Aloha , Ore. , \$ 5,000 and six-month suspension ; Mona Sun , Jamaica Estates , N.Y. , \$ 60,000 ; William Swearingen , Minneapolis , \$ 15,000 and six-month suspension ; John Bew Wong , San Francisco , \$ 25,000 ; Rabia M. Zayed , San Francisco , \$ 50,000 .	†
18	Andrew Derel Adams , Killeen , Texas , fined \$ 15,000 ; John Francis Angier Jr. , Reddington Shores , Fla. , \$ 15,000 ; Mark Anthony , Arlington Heights , Ill. , \$ 10,000 and 30-day suspension ; William Stirlen , Arlington Heights , Ill. , \$ 7,500 and 30-day suspension ; Fred W. Bonnell , Boulder , Colo. , \$ 2,500 and six-month suspension ; Michael J. Boorse , Horsham , Pa. ; David Chiodo , Dallas , \$ 5,000 , barred as a principal ; Camille Chafic Cotran , London , \$ 25,000 ; John William Curry , fined \$ 5,000 , ordered to disgorge \$ 30,000 , one-year suspension .	†
16	8.60 % 30 to 44 days ; 8.55 % 45 to 59 days ; 8.375 % 60 to 79 days ; 8.50 % 80 to 89 days ; 8.25 % 90 to 119 days ; 8.125 % 120 to 149 days ; 8 % 150 to 179 days ; 7.625 % 180 to 270 days .	DM
14	Among other OTC issues , Intel , dropped 2 18 to 33 78 ; Laidlaw Transportation lost 1 18 to 19 12 ; the American depositary receipts of Jaguar were off 14 to 10 14 ; MCI Communications slipped 2 14 to 43 12 ; Apple Computer fell 3 to 45 34 and Nike dropped 2 14 to 66 34 .	
<i>continued on the next page</i>		

<i>continued from the previous page</i>		
12	8 1316 % to 8 1116 % one month ; 8 1316 % to 8 1116 % two months ; 8 1316 % to 8 1116 % three months ; 8 34 % to 8 58 % four months ; 8 1116 % to 8 916 % five months ; 8 58 % to 8 12 % six months .	D
12	8.52 % 30 days ; 8.37 % 60 days ; 8.15 % 90 days ; 7.98 % 120 days ; 7.92 % 150 days ; 7.80 % 180 days .	DM
12	“ Fundamentally dangerous ... , ” Mr. Smith said , almost in a whisper , “ fundamentally weak ... fairly vulnerable still ... extremely dangerously poised ...	
11	“ There may be sticker-shock reaction initially , ” said Mr. Pratt , “ but as the wine is talked about and starts to sell , they eventually get excited and decide it ’s worth the astronomical price to add it to their collection . ”	
11	Thus , in a civil case , a defendant may be called as a witness , he may be forced to testify or take the Fifth , and his taking of the Fifth may permit the drawing of an adverse inference against him in the civil matter .	M
11	“ They ’ve been laggard , ” he says , “ but they ’ll have to become more aggressive . ”	DM
11	“ To get people ’s attention these days , ” says Douglas Bailey , a political consultant , “ your TV ad needs to be bold and entertaining , and , more often than not , that means confrontational .	? ?
10	If takeover premiums become excessive , if LBO dealmakers become too aggressive , then the private market will recognize these problems more quickly and accurately than will policy makers , and the markets will move with lightning speed to impose appropriate sanctions .	
10	“ One of the things that continues to worry me is this monetary warfare between the Treasury Department and the Federal Reserve Board , ” said Lawrence Kudlow , a Bear , Stearns & Co. economist , on ABC ’s “ This Week . ”	DM
10	Canada 13.50 % ; Germany 8.50 % ; Japan 4.875 % ; Switzerland 8.50 % ; Britain 15 % .	DM

Table 1: The most difficult sentences selected with respect to the number of top-level constituents were 10–18 nodes wide.

29	They hope the foreign deals will divide the Hollywood opposition and prod Congress to push for ending federal rules that prohibit the networks from grabbing a piece of rerun sales and owning part of the shows they put on the air .
28	Mr. Otradovec said Boeing told America West that the 757 it was supposed to get this Thursday would n't be delivered until Nov. 7 – the day after the airline had been planning to initiate service at Houston with four daily flights , including three nonstops to Phoenix and one nonstop to Las Vegas .
28	Lawmakers and administration officials agree that Friday 's drop , by itself , is n't enough to force both sides back to the table to try to reach a deficit-reduction agreement that would be more serious and more far-reaching than last spring 's gimmick-ridden plan , which still is n't fully implemented .
28	There is also speculation that Mr. Newhouse could bring in a powerhouse businessman or another Newhouse family member to run the business side , in combination with a publishing executive like Robert Gottlieb , who left Random House 's Alfred A. Knopf to run the New Yorker , also owned by the Newhouse family .
27	But the airlines are scarcely a clear case , given anti-takeover mischief by Secretary of Transportation Skinner , who professes to believe safety will be compromised if KLM and British Airways own interests in companies that fly airplanes .
27	The limits to legal absurdity stretched another notch this week when the Supreme Court refused to hear an appeal from a case that says corporate defendants must pay damages even after proving that they could not possibly have caused the harm .
27	Perhaps none of the unconstitutional conditions contained in the appropriations bills for fiscal 1990 better illustrates Congress 's attempt to usurp executive power than Section 609 of the executive-office bill : “ None of the funds made available pursuant to the provisions of this Act shall be used to implement , administer , or enforce any regulation which has been disapproved pursuant to a resolution of disapproval duly adopted in accordance with the applicable law of the United States . ”
<i>continued on the next page</i>	

<i>continued from the previous page</i>		
27	An Upjohn spokesman said he had “ heard nothing ” to suggest the early retirement package was spurred by shareholder pressure or a potential bidder for the company , which occasionally has been the target of takeover speculation .	M
27	Despite the harsh exchanges , the U.S. and China still seem to be looking for a way to mend relations , which have deteriorated into what Mr. Nixon referred to as “ the greatest crisis in Chinese-American relations ” since his initial visit to China 17 years ago .	
26	But he is best known in the auto industry as the creator of a team car-development approach that produced the two mid-sized cars that were instrumental in helping the No. 2 auto maker record profits in recent years and in enabling the company ’s Ford division to eclipse General Motors Corp. ’s Chevrolet division as the top-selling nameplate in the U.S. .	
26	The problem , however , is that GM ’s moves are coming at a time when UAW leaders are trying to silence dissidents who charge the union is too passive in the face of GM layoffs .	d
26	An NBC spokesman counters that Mr. Holston ’s lament was “ entirely consistent ” with NBC plans because the U.S. rules would limit NBC ’s involvement in the Qintex deal so severely as to be “ light years away from the type of unrestrained deals available to Sony – and everyone else except the three networks . ”	
26	The exchange said it feared that some members would n’t be able to find enough soybeans to deliver and would have to default on their contractual obligation to the Italian conglomerate , which had refused requests to reduce its holdings .	
26	Another proposed reform is to have program traders answer to an “ uptick rule ” a reform instituted after the Great Crash of 1929 that protects against stocks being relentlessly beaten downward by those seeking to profit from lower prices , namely short sellers .	
<i>continued on the next page</i>		

<i>continued from the previous page</i>		
25	Rather , it is born of frustration with having to combat constantly changing strains of a statist idea that one thought had been eliminated in the early 1970s , along with smallpox .	

Table 2: The most difficult sentences selected with respect to the tree height were 25–29 nodes high.

53	GOODY PRODUCTS Inc. cut its quarterly dividend to five cents a share from 11.5 cents a share .	d
53	In investing on the basis of future transactions , a role often performed by merchant banks , trading companies can cut through the logjam that small-company owners often face with their local commercial banks .	
53	The Fed cut the key federal funds interest rate by about 0.25 percentage point to 8.75 % after the Oct. 13 stock market plunge , but has shown no sign of movement since .	
53	In the hands of a zealot like Lenny Bruce , this double-edged blade could cut both the self and the audience to ribbons .	D
53	Congress previously cut six airports this year .	DM
53	As a part of overall efforts to reduce spending , Congress cut by \$ 30 million the Bush administration 's request for antitrust enforcement for fiscal 1990 , which began Oct. 1 .	M
53	Magna recently cut its quarterly dividend in half and the company 's Class A shares are wallowing far below their 52-week high of 16.125 Canadian dollars (US\$ 13.73) .	d M
51	Also , retail sales grew 0.5 % last month .	D
51	And the nose on Mr. Courter 's face grows .	DM
51	Per-capita personal income in the U.S. grew faster than inflation last year , according to the Bureau of Economic Analysis .	D
51	The gap between winners and laggards will grow .	DM
51	Producer prices for intermediate goods grew 0.4 % in September , after dropping for three consecutive months .	d
51	The company grew modestly until 1986 , when a majority position in Hooker Corp. was acquired by Australian developer George Herscu , currently Hooker 's chairman .	D
51	Sales grew almost 7 % to \$ 279.1 million from \$ 261.3 million .	D
51	Sales of passenger cars grew 22 % from a year earlier to 361,376 units .	d

Table 3: The most difficult sentences selected with respect to the top-level verb ambiguity score in the observed corpus had either *cut* (7 sentences) or *grow* (8 sentences) as the top-level verb.

References

- [MS] C.D. Manning and H. Schütze, “Foundations of statistical natural language processing”, MIT Press, 1999.
- [PT] Marcus et al., “Building a large annotated corpus of English: the Penn Treebank”, Computational Linguistics, Vol 19., 1993,
- [DL] D. Lin, “Dependency-based Evaluation of MINIPAR”. In Workshop on the Evaluation of Parsing Systems, Granada, Spain, May 1998.
- [DB] D. Bikel. “Intricacies of Collins’ parsing model”, Computational Linguistics, 2004.
- [UG] D. Bikel. “dbparser User Guide”, <http://www.cis.upenn.edu/~dbikel/download/dbparser/guide.pdf>, September 2005.
- [YG] Yoav Goldberg, Bikel’s parser coordination results on a subset of the Brown corpus, drafts, December 2005.