Algorithms for AI and NLP (Fall 2011, Assignment 2b)

Goals

1. Inspect and understand the inner workings of our bottom-up left-corner parser;
2. adapt the generalized chart parser to one-best probabilistic unpacking from the forest;
3. implement the ParsEval metrics and evaluate quantitatively the performance of the parser.

1 Estimating a PCFG from the Penn Treebank

For this session, our course set-up provides the model solution to the previous assignment as a starting package, called '`pcfg.lsp`'. Please revise and extend the code in this file.

Even though our chart parser can construct complete parse forests for sentences up to some non-trivial length in reasonable time and memory, seeing the massive ambiguities admitted by our PTB-derived grammar (see below), exhaustive unpacking is inconceivable for all but the shortest parser inputs. Fortunately, typical applications of syntactic parsing are most interested in the most probable analysis (rather than in millions of possible analyses). In this part of the problem set, we will finally put to use the rule and lexeme counts that we worked so hard to acquire from the PTB files earlier.

(a) Augment our function `read-grammar()` to determine rule and lexeme probabilies, using the standard maximum likelihood estimation 'idiom' for conditional probabilities. To avoid costly multiplications (using expensive rational numbers), transpose the probabilities into logarithmic space (so-called *log-probabilities*), as discussed in the lecture. Recall that the following is true: $log\prod_i P_i = \sum_i \log P_i$.

(b) While we are at post-processing the raw grammar (as read from '`wsj.mrg`'), observe that a relatively large proportion of rules is seen with very low counts only. As there is reason to be mildly sceptical of such rules (which might reflect annotation inconsistencies in the training data, rather than true linguistic generalizations) as well as of our ability to estimate reliable probabilities for them, implement a *frequency threshold* on grammar rules: for some value n, we will choose to discard all rules that were observed less than n times in training. To get started, set n to a conservative value of 1, but once we have a complete parsing and evaluation pipeline, feel free to experiment with more aggressive frequency thresholding.

(c) Now that our grammar is a little smaller and all rules and lexemes have probabilities attached to them, we will need to augment the edge structure to keep track of the probabilities of entries in our chart. During forest construction, we will use the *probability* slot in the edge structures to record the probability of the underlying part of the grammar, i.e. the rule or lexeme that licensed each edge; these 'local' probabilities on each node of the packed forest will later be used in a one-best unpacking phase, to compute probabilities of full sub-trees, and for each packed edge to find the sub-tree among all the trees represented by that edge that has the maximum probability. Go through the core functions of the chart parser and make sure that all newly created edges have their probabilities set properly. For the corner case of edges created in chart initialization for the actual input tokens (i.e. the daughter edges to lexemes), assume that they are certainties, in other words each should have the largest possible probability.

2 Bottom-Up Left-Corner Parsing

(a) In our starting package, we have made the necessary changes for the chart parser to work with the revised representation of the grammar. Thus, in principle the code is ready to parse sentences using the large, PTB-derived grammar; however, there are some scalability issues.

(a) To confirm the basic parsing functionality, train the grammar from the '`wsj.mrg`' file and then try gently (i.e. with a short sentence), for example:

```
? (setf edges (parse '("It" "was" "Monday" ".")))
? (unpack-edge (first edges))
```

With a little bit of patience, the above example calls should complete and show meaningful results. Observe the total number of parse trees for this example; given our intuitions (as near-native speakers) about English, is this a plausible number of distinct syntactic analyses? Assuming the code in our starting package is correct, how can we explain what happens in parsing this very simple sentence using a grammar read off the annotations in the PTB?

(b) For testing purposes, our course set-up provides an additional file ‘`test.mrg`’ (Section 23) of the Penn Treebank, which is the commonly used test data. Although we may not quite be able to make our parser process all sentences from this file (on commodity hardware), the `parse()` function at least is sufficiently efficient to be called on any sentence from ‘`test.mrg`’ of up to twenty tokens in length. Write a function to extract the sequence of leaves (i.e. a list of strings) from the ‘gold-standard’ trees in ‘`test.mrg`’ and run those below twenty tokens through the parser. Without unpacking, count the number of edges returned from the `parse()` function; by and large, you should only observe two different counts. Explain why this makes sense.

(c) To better understand the internals of the edge structure, write a function `edge-to-tree()` that takes one of the edges returned by `parse()` as its input and returns the full parse tree of that edge (ignoring all local ambiguity packing) as a Lisp list, i.e. the same format used in the PTB files. For inspiration in writing this function, please observe how `unpack-edge()` recursively traverses the edge structure.

3 Probabilistic One-Best Unpacking

(a) To adapt the Viterbi algorithm to the extraction of the most probable complete tree from a parse forest, we need to revisit our interpretation of the *probability* slot of the edges in the forest, once forest construction is complete. In this part of the problem set, our goal is to develop a new function that operates on an edge returned by `parse()`, where typically there will be packed local ambiguity at many levels. Replacing our predefined exhaustive (and computationally not tractable) procedure `unpack-edge()`, this new function will recursively traverse both daughter and ‘alternate’ (i.e. packed) edges and for each node find the sub-tree with the maximum probability for the corresponding sub-string and category assignment. Ignoring ambiguity packing for the time being, the probability of an edge in Viterbi unpacking should reflect the probability of the tree represented by that edge, i.e. (conceptually) the product of probabilities of all lexemes and rules contributing to the tree. To get started, write a function `viterbi()` that recomputes probability values for all the ‘primary’ edges, i.e. the top-level result and all its daughters (but not alternates), to now reflect full sub-tree probabilities; the new values should be recorded in each node of the tree (of edges).

(b) Now generalize the Viterbi algorithm to the parse forest, i.e. to take into account the disjunctive nodes corresponding to local ambiguity packing. Given all information packed into our edges and the parse forest, we are able to read off the most probable tree in a single pass, without additional search. Observe that forest construction effectively is the equivalent to filling in the HMM trellis, such that the one addition to the management of per-edge probabilities (from part (a) above) is to decide on what to do in the face of ambiguity packing. Much like in the Viterbi algorithm for lattices (i.e. the HMM case), there will need to be a maximization step at each node where there are disjunctive alternative. However, unlike in the HMM case, there should be no need for auxiliary data structures, as we can destructively modify the relevant parts of each edge structure. In other words, the complete `viterbi()` function should have two side effects on each edge, viz. (a) recording the probability of the highest-scoring sub-tree represented by this edge and (b) adjusting the list of daughters to the sequence of daughter edges that gave rise to that very sub-tree. Decide whether there is any need for dynamic programming in this ‘trivial’ (i.e. one-best) unpacking procedure; if so, consider reusing the *cache* slot that is present already in our edge structure.

4 Evaluating our Parser

To get a better idea of how well our parser actually performs (on unseen inputs, i.e. sentences not contained in the training data), implement the ParsEval metric (please see slide #17 from our most recent lecture). To compute ParsEval scores, it might be convenient to have an auxiliary function that decomposes one tree into a set of labelled bracketings, where each bracketing $\langle C, i, j\rangle$ captures the syntactic category C assigned to the sub-string of input starting at i and ending at j. Note that it is customary to *not* include PoS tags (i.e. the preterminal nodes of the tree, or categories of lexemes) in ParsEval scores. In case it seems wasteful to you to explicitly enumerate two sets of bracketings only to count overlapping and non-overlapping elements, consider an alternate scheme of computing, for a pair of trees, the relevant counts.

(a) Parse all sentences of Section 23 of the PTB (i.e. the leafs of the trees in our file '`test.mrg`') using our chart parser, extract the one-best Viterbi tree, and compute ParsEval scores against the gold-standard trees. Recall that the combined F_1 measure in ParsEval is the harmonic mean of precision and recall of labelled bracketings. Owed to unknown words in the test data, our parser will likely fail to parse some sentences; what should be the contribution of these test inputs to the overall ParsEval scores?

(b) To put parser evaluation figures into perspective, it may be useful to construct a so-called *baseline*, i.e. a score reflecting what would be the result without the Viterbi step. Construct a naïve variant of one-best unpacking that effectively ignores all probabilities, i.e. simply returns the tree that the parser happened to find first. Does our probabilistic chart parser improve over this baseline in terms of ParsEval scores?

5 Towards a More Realistic Treebank Parser (Optional)

There are a number of common optimizations in statistical parsing that our current implementation lacks. It is actually not uncommon for a state-of-the-art parser trained on the PTB to use about one second for parsing a (comparatively complex) input, but nevertheless our current parser probably is one or two order of magnitudes less efficient than cutting-edge statistical parsers. Furthermore, we still fail to provide any syntactic analysis when there is a single unknown word in our input, where unknown tokens include, for example, all names and numbers not observed in training.

(a) To improve robustness to unknown words, combine the chart parser with our HMM tagger. Recall that the tagger was trained on the exact same data and achieved a tagging accuracy of around 96 % on the unseen test data from Section 23. One could imagine at least two different ways of using the tagger to preprocess parser inputs: (a) for unknown words, i.e. input tokens for which there is no lexical entry in our grammar, the PoS of the token could be provided by the tagger; this should effectively make it possible to parse all sentences (within a suitable upper bound on input length, as before) from '`test.mrg`'; (b) to further improve parser efficiency, the tagger could be used to reduce lexical ambiguity, where lexical look-up in the grammar would effectively be replaced with the PoS sequence obtained from most probable path through the HMM. Experiment with both methods of coupling the tagger and parser, and report on your experimental findings.

(b) Another technique to improve the efficiency of the parser is so-called *chart pruning*. The basic idea is, for each chart cell, to discard partial analyses with very low relative probabilities. Typically, this is accomplished by assuming a cut-off beam θ, where for each cell edges whose probability is less than $1/\theta$ of the best-scoring edge in that same cell are discarded. Assume that, at some point during forest construction, for cell $\langle i, j \rangle$ there is an edge e_1 whose probability p_1 is the highest probability for all edges in that cell. The basic intuition in chart pruning is that a new edge e_2 in that same chart cell, with a probability $p_2 < 1/\theta p_1$, is very unlikely to give rise to a tree whose total probability is larger than any tree built using e_1. Think about the assumptions we are making here, and explain why it is possible in principle that a tree containing e_2 might end up with a higher probability than all trees containing e_1.

If you have not done so already, rework the global chart as an abstract data type providing the various types of efficient indexing we need for our chart parser. Revise the probability computations during forest construction, add a suitable level of accounting of per-cell maximum probabilities, and then experiment with chart pruning and various levels of θ.

6 Submitting Your Results

Please submit your results in email to Johan ('`johanbev@ifi.uio.no`'), Jonathon ('`jread@ifi.uio.no`'), and Stephan ('`oe@ifi.uio.no`') before the end of the day on Thursday, December 1. Please provide all files that you extended (or created) as part of this exercise (e.g. minimally '`pcfg.lsp`'), including all code and answers to the questions above.