Extracting tree fragments in linear average time

Andreas van Cranenburgh
acranenb@science.uva.nl

July 12, 2012

Abstract

This report details the implementation of a fragment extraction algorithm using an average case linear time tree kernel. Given a treebank, the algorithm extracts all fragments that occur at least twice, along with their frequency. Evaluation shows a 70-fold speedup over a quadratic fragment extraction implementation. Additionally, we add support for trees with discontinuous constituents.

1 Introduction

Given a collection of tree structures, a useful question concerns the recurring patterns that occur in it. Kernel methods, which quantify similarity in some way, and specifically tree kernel methods, exist to operationalize this question. An additional question is to explicitly enumerate these patterns, so as to obtain not only a numeric value about the similarity of structures, but to find out what it consists of; these objects could be considered its building blocks. In computational linguistics this has applications in corpus linguistics and natural language parsing, and applications in other domains are possible as well.

The notion of a fragment is characterized as follows:

Definition. *A fragment f of a tree T is a connected subset of nodes from T, with $|f| \geq 2$, such that each node of f has either all or none of the children of the corresponding node in T.*

An algorithm to extract recurring fragments returns the largest fragments that occur in any pair of trees in the input. As an example, given these trees as input:

```
(S (NP (DT The) (NN cat)) (VP (VBP saw) (NP (DT the)
    (NP|<JJ-NN> (JJ hungry) (NN dog)))))
(S (NP (DT The) (NN cat)) (VP (VBP saw) (NP (DT the) (NN dog))))
```

We want the algorithm to find the following maximal common fragments (with frequencies in parentheses):

```
(S (NP (DT The) (NN cat)) (VP (VBP saw) (NP ))) 2
(NP (DT ) (NN ))        3
(DT the)        2
(NN dog)        2
```

An algorithm for the extraction of recurring fragments was first presented by Sangati et al. (2010). That algorithm compares each node in the input to all others, giving a best and worst-case quadratic complexity (i.e., $\Theta(n^2)$). A fast tree kernel was presented by Moschitti (2006), with an average case linear complexity (i.e., $O(n)$ on average). However, his version only returns a list of matching nodes. This work presents an implementation of the fast tree kernel that extracts recurring fragments, providing a significant speedup.

2 Iterating over the treebank

Each tree is represented as a list of nodes sorted according to its productions. For each pair of trees, a bit matrix is constructed where the bit at (n, m) is set iff the nodes at those indices in the respective trees have the same production. From this table the bit sets corresponding to fragments are collected and stored in the results table. We generalize the task of finding recurring fragments in a treebank to the task of finding the common fragments of two, possibly equal, treebanks. In case the treebanks are equal, only half of the possible tree pairs have to be considered: the fragments extracted from $\langle t_n, t_m \rangle$, with $n < m$, are equal to those of $\langle t_m, t_n \rangle$.

The code in this report is in a superset of the Python language that includes type declarations; this allows translation of the code to C using Cython (Behnel et al., 2011).

```
def extractfragments(trees1, trees2):
    results = {}
    SLOTS = MAXNODE / (sizeof(ULong) * 8) + 1
    CST = bitmatrix(MAXNODE * MAXNODE) #Common Subtree Table
    for a in trees1:
        for b in trees2:
            # initialize table
            memset(CST, 0, b.len * SLOTS * sizeof(ULong))
            # fill table
            fasttreekernel(a.nodes, b.nodes, CST)
            # extract results
           getfragments(CST, a.nodes, b.nodes, b.root, results, SLOTS)
    return results
```

Some implementation details: we use arrays of `unsigned long` (abbreviated `ULong`) to store bit-vectors and bit-matrices. The trees are represented as arrays of `Node`s, which contain indices for their left and right descendants.[1] The index `-1` is used as a sentinel value to indicate that there is no production or left node (terminals), or no right node (unary productions). This does mean that the representation requires binary trees, but this simplifies the algorithms

1 Instead of array indices direct pointers could have been used, but these take four times as much memory: 8 bytes on a 64 bit machine, versus 2 bytes for a `short int`.

considerably and most statistical parsers rely on binary trees anyway. Since the bit-vectors and -matrices are dynamically allocated, we pass around pointers and perform manual indexing (i.e., multi-dimensional arrays are simulated by computing indices on a one-dimensional array).

3 The fast tree kernel

The insight that makes this kernel fast on average is that it can be viewed as the problem of finding the intersection of two sequences that have been sorted in advance. In our implementation the productions of each node is mapped to an integer, so that comparisons are cheap. We sort in descending order, so that lexical nodes which have a sentinel production of -1 end up in the tail.

The following code is a direct implementation of the pseudo-code in Moschitti (2006):

```
cdef void fasttreekernel(Node *a, Node *b, ULong *CST, int SLOTS):
    cdef int i = 0, j = 0, jj = 0
    while a[i].prod != -1 and b[j].prod != -1:
        if   a[i].prod < b[j].prod: j += 1
        elif a[i].prod > b[j].prod: i += 1
        else:
            while a[i].prod == b[j].prod:
                jj = j
                while a[i].prod == b[jj].prod:
                    SETBIT(&CST[jj * SLOTS], i)
                    jj += 1
                i += 1
```

Given the trees from the introduction as input, the matrix in table 1 obtains. The matrix visualizes why the algorithm is efficient: there is a path along which comparisons have to be made, but most node pairs (i.e., the ones with a different label on the left-hand side) do not have to be considered. The more productions, the higher the efficiency. In case there is only a single non-terminal label X, and hence only one phrasal, binary production, the efficiency of the algorithm disappears and the worst-case quadratic complexity will result.

4 Extracting maximal connected subsets

After the matrix with matching nodes has been filled, we need to extract nodes belonging to each maximal fragment. A fragment is a connected subset of nodes. We traverse the second tree in depth-first order, in search for possible root nodes of fragments.

To scan the bits of a row of the matrix we use the function `nextset`, which returns the next 1-bit starting from a specific index. This function exploits a CPU instruction which scans for the next 1-bit in a word sized chunk (typically 64 bits). While the extraction of bit sets technically has quadratic time complexity because we walk through a 2-dimensional matrix looking for 1-bits, in practice the bit operations are $O(1)$, not linear, given that maximum number of nodes in a tree from the treebank will be a small, constant multiple of the machine word size. During extraction, only matching nodes are considered,

	0	1	2	3	4	5	6	7	8
	VP	VBP	S	NP	NP	NN	NN	DT	DT
0	VP	1							
1	VBP		1						
2	S			1					
3	NP|<JJ-NN>								
4	NP								
5	NP				1	1			
6	NN						1		
7	NN							1	
8	JJ								
9	DT								1
10	DT								1

Table 1: Matrix of two compared trees.

which means that the same average time complexity obtains as in the algorithm of Moschitti (2006).

```
cdef void getfragments(ULong *CST, ULong *scratch, Node *a, Node *b,
    short j, dict results, int SLOTS):
    cdef short i
    if j < 0 or b[j].prod < 0: return
    while True:
        i = nextset(&CST[j * SLOTS], 0, SLOTS)
        if i == -1: break
        memset(scratch, 0, SLOTS * sizeof(ULong))
        extractat(CST, scratch, a, b, i, j, SLOTS)
        results[getfragment(tree, scratch)] += 1
    getfragments(CST, scratch, a, b, b[j].left, results, SLOTS)
    getfragments(CST, scratch, a, b, b[j].right, results, SLOTS)
```

Whenever a 1-bit is encountered, both trees are traversed in parallel from that node onwards, to collect the nodes. Both trees need to be considered to ensure that extracted subsets are connected in both trees.

```
cdef void extractat(ULong *CST, ULong *result, Node *a, Node *b,
    short i, short j, int SLOTS):
    SETBIT(result, i)
    CLEARBIT(&CST[j * SLOTS], i)
    if a[i].left < 0: return
    if TESTBIT(&CST[b[j].left * SLOTS], a[i].left):
        extractat(CST, result, a, b, a[i].left, b[j].left, SLOTS)
    if a[i].right < 0: return
    if TESTBIT(&CST[b[j].right * SLOTS], a[i].right):
        extractat(CST, result, a, b, a[i].right, b[j].right, SLOTS)
```

Here is an example which demonstrates why fragments must be extracted from both trees in parallel:

```
(S (A (B x) (S y)) (B p))
(A (B x) (S (A y) (B z)))
```

These trees have two productions in common; however, these do not form a contiguous fragment in both trees, because the productions appear in a different order in the respective trees.

Another concern is whether extracting fragments from a pair of trees is a commutative operation; i.e., whether the order of the operands has an effect on the output. Consider the following corpus:

```
(TOP (S (A x)) (S (A b)))
(TOP (S (A x)))
```

Using this order, we get two fragments with the `FragmentSeeker` of Sangati et al. (2010):

```
(S (A "x"))     1
(S A)   1
```

But when the order of the input is reversed, the second fragment is not extracted, because it is a subset of the first. With our algorithm, the output is the same regardless of the order of the input; two fragments are extracted from this example.

5 Discontinuity

The treatment of trees with discontinuous constituents is straightforward, requiring no modification of the extraction algorithm itself. Nodes and their children are traversed as usual, while the context-free productions at each node are replaced by productions of a Linear Context-Free Rewriting System (LCFRS), a formalism which can express discontinuity (Maier and Søgaard, 2008). Such productions distinguish the possible ways spans of children can be combined to form the parent node. The output will contain the tree-structure of the fragments with indices instead of nodes, and the terminals specified separately in a space separated list. For example:

```
(AP (PP 0) (AP|<ADV-ADJD> (ADV 2) (ADJD 3)))    Für  gerade recht   18
```

This fragment was extracted from the German Negra corpus (Skut et al., 1997). The fragment has a gap (discontinuity) between *Für* and *gerade*, which is why there are two spaces between these words.

6 Correctness & efficiency evaluation

The work performed by the algorithm can be efficiently distributed among the available cores. To abstract over the number of available cores, we not only report the wall clock time but also the total CPU time used by all cores, which is comparable to the time that would have been spent if a single core was available. As treebank we use the Wall Street Journal WSJ section of the Penn treebank (Marcus et al., 1993). In a pre-processing step we binarize the training section (2–21) of the treebank with $h = 1$, $v = 2$ markovization (i.e.,

horizontal context limited to a single sibling, one vertical parent), left-factored. To demonstrate the capability of working with discontinuous treebanks, we use the German Negra treebank (Skut et al., 1997), binarized $h = \infty$, $v = 1$.

Aside from the fast tree kernel just discussed, we also test with a re-implementation of the quadratic tree kernel, to compare the effects of the choice of programming language and representation. The results are in table 2.

	Time (hr:min)		
Implementation	CPU	Wall clock	fragments
Sangati et al. (2010):			
Quadratic tree kernel, WSJ	160	10:00	1,023,092
This work:			
Quadratic tree kernel, WSJ	93	6:15	1,032,568
Fast tree kernel, WSJ	2.3	0:09	1,023,880
Fast tree kernel, Negra	0.8	0:04	370,081

Table 2: Performance comparison. Wall clock time is when using 16 cores.

Our quadratic tree kernel gets a modest 1.7 fold speedup, which is probably due to the more low-level style of programming. The real gain comes from the asymptotic speedup of the fast tree kernel: we obtain a 70-fold speedup over `FragmentSeeker`; 40-fold over our quadratic tree kernel.

The results show slight variation in the number of fragments found. To validate our results, we compare the output of our system to that of `FragmentSeeker`. If the output of the latter is taken as a gold standard, and we disregard frequencies, we get an F_1 score of 99.93 % for our implementation of the fast tree kernel; i.e., there are only about 1500 fragments over which there is disagreement.

If exact frequencies[2] are computed the fast tree kernel needs 13 minutes; the frequencies match exactly with those of `FragmentSeeker`. With `FragmentSeeker` exact frequencies require another 5 hours to compute. For each fragment, the whole treebank is traversed to count its occurrences, which is again quadratic but with larger constant factors due to the number of nodes in recurring fragments. We need only 4 minutes by employing an index of the sets of trees containing a particular production. By taking the intersection of the sets for the productions in a fragment, we get a precise list of candidate trees which could contain the fragment one or more times. The actual number of occurrences is then counted by traversing these trees.

7 Applications

A set of recurring fragments can be converted into a Data-Oriented Parsing (DOP) grammar—a method called Double-DOP (Sangati and Zuidema, 2011). Using our fragment extractor, we can apply this method to treebanks with discontinuous constituents as well. Another application is to use the fragments to define a similarity measure between texts, which can be applied to the task of authorship attribution (van Cranenburgh, 2012). Lastly, any kind of labelled,

2 Approximate frequencies are returned by default, which are the frequencies of extracting a fragment as maximal fragment of a pair of trees, while exact frequencies include all occurrences.

tree-structured data set could be analyzed in terms of its patterns using the algorithm presented here.

8 Conclusion

We have presented an implementation of a fragment extraction algorithm using an average case linear time tree kernel. We obtain a substantial speedup over the previously presented quadratic algorithm, and the resulting fragments and frequencies have been validated against the output of the latter. Additionally, we introduced support for discontinuous constituents.

The source code of our implementation is available for download as part of `disco-dop`, cf. `https://github.com/andreasvc/disco-dop`

Acknowledgements

Thanks to Federico Sangati for making the source code of his `FragmentSeeker` available, and Rens Bod for reading a draft of this report.

References

Behnel, Stefan, Robert Bradshaw, Craig Citro, Lisandro Dalcin, Dag Sverre Seljebotn, and Kurt Smith (2011). Cython: The best of both worlds. *Computing in Science and Engineering*, 13:31–39.

Maier, Wolfgang and Anders Søgaard (2008). Treebanks and mild context-sensitivity. In *Proceedings of Formal Grammar 2008*, page 61.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Moschitti, Alessandro (2006). Making tree kernels practical for natural language learning. In *Proceedings of EACL*, pages 113–120. Available from: `http://acl.ldc.upenn.edu/E/E06/E06-1015.pdf`.

Sangati, Federico and Willem Zuidema (2011). Accurate parsing with compact tree-substitution grammars: Double-DOP. In *Proceedings of EMNLP*, pages 84–95. Available from: `http://www.aclweb.org/anthology/D11-1008`.

Sangati, Federico, Willem Zuidema, and Rens Bod (2010). Efficiently extract recurring tree fragments from large treebanks. In *Proceedings of LREC*, pages 219–226. Available from: `http://dare.uva.nl/record/371504`.

Skut, Wojciech, Brigitte Krenn, Thorten Brants, and Hans Uszkoreit (1997). An annotation scheme for free word order languages. In *Proceedings of ANLP*, pages 88–95.

van Cranenburgh, Andreas (2012). Literary authorship attribution with phrase-structure fragments. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 59–63, Montréal, Canada. Available from: `http://www.aclweb.org/anthology/W12-2508`.