CSE 254 (Spring 2003)

"Growing N-gram Trees for Language Modeling"

by Dustin Boswell

June 7, 2003

Language Modeling

- Given a stream of words, we want to predict the next word.
- Ex: `... the president was able to _____`
- That is, given a history h, we want a distribution $P(w|h)$.
- We estimate using statistics from a training text.
- `P( conference | machine learning ) =`

$$\frac{\texttt{count(machine learning conference)}}{\texttt{count(machine learning)}}$$

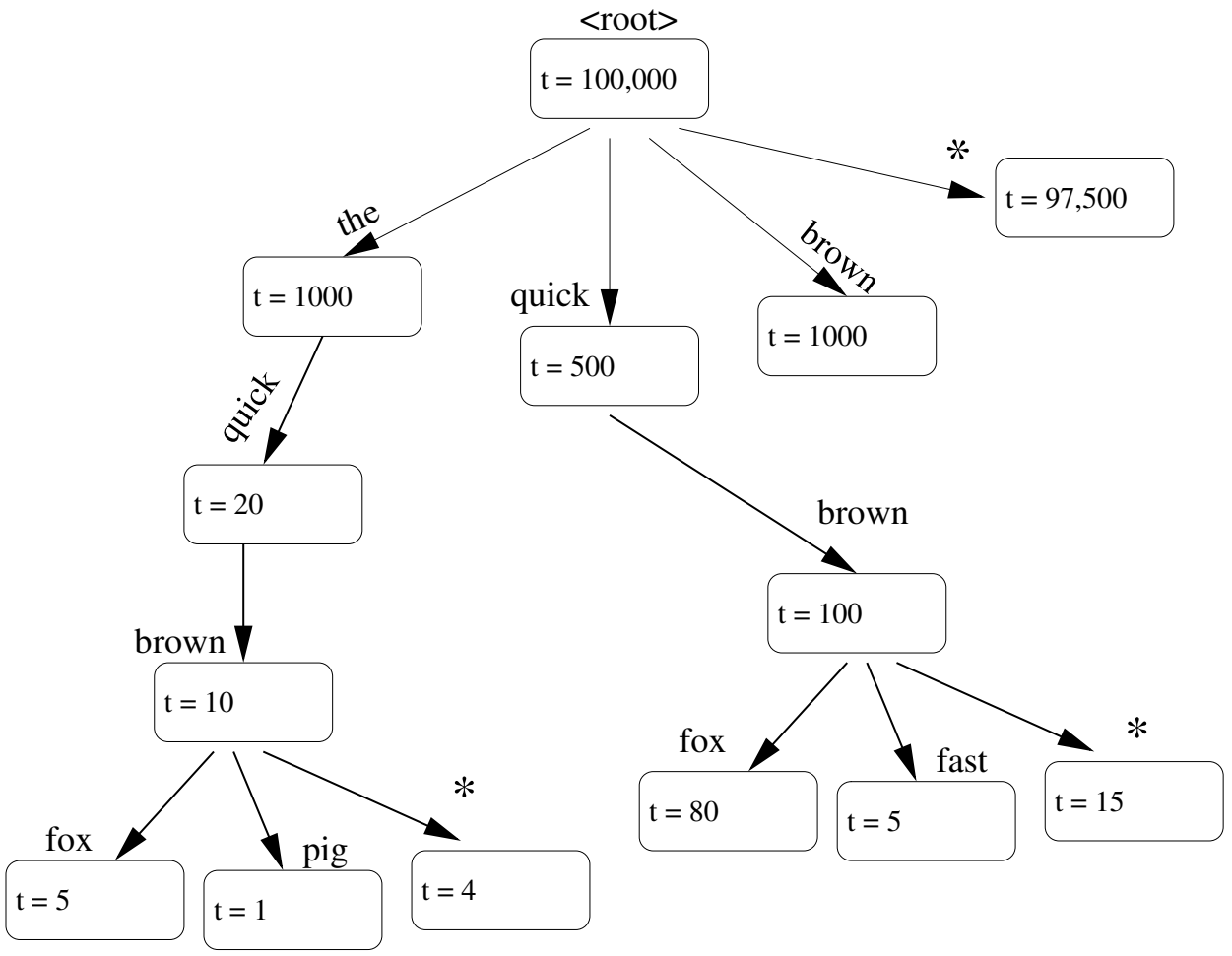
N-gram model

- Choose a fixed $N = 2$ or 3.

- Obtain all statistics for $|h| = N - 1$

- Problems
  - Space requirements $\approx |Vocab|^N$.
    (Ex: $20,000^3 \gg MEMORY\_SIZE$)
  - What if histories of size $> N - 1$ are important?

Variable N-Grams Trees



Tree-Growing Algorithm

1 Start with an empty tree.

2 Scan through training corpus.
(Keeping track of which histories come up)

3 Add the most requested leaves to the tree.

4 Repeat steps 2-3 as memory permits.

Portions of the actual Word Trie generated

```
NODE  traffic  child_traffic  num_children
-----------------------------------------
<ROOT> 53771683 53156023 32245
        ...
        across 8514 6907 23
                america 114 0 0
                ...
                the 5310 3268 20
                        board 388 145 2
                                COMMA 89 0 0
                                PERIOD 56 0 0
                        ...
                        street 284 198 3
                                COMMA 37 0 0
                                PERIOD 43 0 0
                                from 118 77 1
                                        the 77 0 0
```

Portions of the actual Word Trie generated

```
NODE  traffic  child_traffic  num_children
-----------------------------------------

       ...
       adjusted 1613 887 7
                annual 271 235 1
                       rate 235 216 1
                              of 216 67 1
                                     1 67 55 1
                                            POINT 55 0 0


       ...
       of 1198167 1025437 2945
                1 24883 24749 16
                       POINT 2455 2455 10
```

Perplexity Results of WordTrie(250,000 nodes) vs. CMU(trigram)

		testing set →			
	training set ↓	sample3	sample4	sample5	NY96
CMU	sample3	131	310		
	sample4		246	284	
	sample5			263	292
WT	sample3	109	314		
	sample4		233	278	
	sample5			274	305