# Syntactic Tokens

AMS Special Session on
Sequences, Words, and Automata
Joint Mathematics Meetings
Wednesday, January 15, 2020 10:00
Room 403 Colorado Convention Center
Denver, CO

Andrzej Ehrenfeucht
andrzej@cs.colorado.edu

# Introduction

Replacing a sequence of characters by a sequence of short strings called “tokens” is often the first step in the analysis of a text. Tokens can be strings belonging to the lowest grammatical categories, or they can be the smallest meaningful units called morphemes. Tokens allow us to construct a grammar independent of the “surface structure” of a language that includes features of the specific alphabet in which the language may be written.

We call “syntactic tokens” short words written in a specific alphabet A, selected by their structure. Such tokens can be used to compare similarities and differences between words from different languages that are written in the same alphabet.

# Informal definition of syntactic tokens

The set of words S over an alphabet A is a set of syntactic tokens if and only if

1) S is finite.
2) No two words in S are sub-words of each other.
3) S is closed under all permutations of letters in A.
4) Every word w $\in$ A* is covered by the occurrences of tokens from S, with the possible exception of a prefix and a suffix of bounded lengths.
   We call the triplet, (uncovered prefix, sequence of covering tokens, uncovered suffix), the description of w, written as des(w).
5) For every v, w $\in$ A*, if des(v) = des(w), then v = w.

A set of syntactic tokens S over alphabet A is maximal, if S is not a proper subset of any other set S' of syntactic tokens over A.

Trivial examples

For every $n > 0$, the set $S_n$ of all words w of length n, $|w| = n$, over alphabet A is a maximal set of syntactic tokens.

The description of the word alibaba, in $S_3$, is

des(alibaba) = ali lib iba bab aba

# Definition of structured words

A word $w \in A^*$ is simple iff every letter $a \in A$ occurs in w at most once.

A word is structured if it is not simple.

A word awa, where $a \in A$ and $w \in A^*$, is a cycle iff the words aw and wa are simple.

Cycles are the smallest structured words. (They are smallest under the relationship: u is a sub-word of w.)

A word awb, where $a, b \in A$ and $w \in A^*$, is bordered iff w is simple and both aw and wb are structured.

Bordered words are the smallest words that contain exactly two cycles.

Examples of sets of all cycles and all bordered words over alphabets

| {a} | | {a, b} | | {a, b, c} | | |
|---|---|---|---|---|---|---|
| aa | aaa | aa | aaa | aa | aaa | 3 |
| | | bb | bbb | bb | abaa | 6 |
| | | aba | aabb | cc | aaba | 6 |
| | | bab | bbaa | aba | aabb | 6 |
| | | | abab | bab | abab | 6 |
| | | | baba | aca | aabca | 6 |
| | | | abaa | cac | babcb | 6 |
| | | | aaba | bcb | cabcc | 6 |
| | | | babb | cbc | aabcb | 6 |
| | | | bbab | abca | aabcc | 6 |
| | | | | acba | babca | 6 |
| | | | | bacb | babcc | 6 |
| | | | | bcab | cabca | 6 |
| | | | | cabc | cabcb | 6 |
| | | | | cbac | total: | 81 |

Sets of all cycles for alphabets {a} and {a, b} are maximal sets of syntactic tokens.

But sets of cycles from bigger alphabets don't satisfy condition (4) above.

The set of bordered words over any alphabet is a maximal set of syntactic tokens.

Example

w = abbacbccabbba

cycles: bb cbc bb

bacb bb

cc

des(w) = a bbacb bacbc cbcc ccabb bbb a

# Number of cycles and number of bordered words

| Cardinality of the alphabet | Number of cycles | Number of bordered words |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 4 | 10 |
| 3 | 15 | 81 |
| 4 | 64 | 625 |
| 5 | 325 | 5,545 |
| 6 | 1,956 | 50,886 |
| 7 | 13,699 | 506,905 |
| 8 | 109,600 | 5.480 E6 |
| 9 | 986,409 | 6.412 E7 |
| 10 | 9.864 E6 | 8.089 E8 |

| Cardinality of the alphabet | Number of cycles | Number of bordered words |
|---|---|---|
| 11 | 1.085 E8 | 1.096 E10 |
| 12 | 1.130 E9 | 1.589 E11 |
| 13 | 1.693 E10 | 2.454 E12 |
| 14 | 2.370 E11 | 4.029 E13 |
| 15 | 3.555 E12 | 7.003 E14 |
| 16 | 5.687 E13 | 1.285 E16 |
| 17 | 9.669 E14 | 2.489 E17 |
| 18 | 1.740 E16 | 5.047 E18 |
| 19 | 3.307 E17 | 1.075 E20 |
| 20 | 6.613 E18 | 2.394 E21 |

# Algorithms for constructing descriptions of words

Let D(A,B) be the set of descriptions des(w) of $w \in A^*$ in terms of the set of the bordered words B.

The des(w) is the function from A* onto B that can by computed by (easy) algorithms having

constant memory complexity: $O(|A|)$

and linear time complexity: $O(|A|*|w|)$

The length of the sequence of words des(w) is

$$|des(w)| \leq |w|$$

When $des(w) = w_1, w_2, .. , w_n$, the sum of the lengths, sl(w), where $sl(w) = |w_1| + |w_2| + \ldots |w_n|$, can be as long as $|A|*|w|$.

But the average length of sl(w), of words of the same length is smaller than or equal to $3*|w|$.

$$average(sl(w)) \leq 3*|w|$$

For example, for $|A| = 20$, $average(sl(w)) \leq 2.265*|w|$.

# Basic syntactic building blocks

There are many other sets of syntactic tokens besides trivial and bordered.

But the large cardinality of sets of syntactic tokens even for a medium sized alphabet and their structural complexity shows that that we need simpler “basic syntactic building blocks” that can be used to describe and compare structures of individual syntactic tokens.

We take all simple words and all cycles as the basic syntactic blocks for words in $A^*$, and define the basic syntactic blocks description of $w \in A^*$ as follows:

bdes(w) is the sequence of all cycles occurring in w, interspersed with all non-empty simple words disjoint from them.

Theorem: If bdes(u) = bdes(w) then u = w.

Example

w = abcacbadefccab bdes(w) = abca cac acba def cc ab

# The structure of bordered syntactic tokens

Let alph(w) be the set of letters from A which occur in w.

If w is a bordered word, then

(1) $|alph(w)| = |w| - 2$

The description bdes(w) contains either two blocks x y or three blocks x z y.

(2) If bdes(w) = x z y, then x and y are cycles, z is simple and their alphabets alph(x), alph(y) and alph(z) are disjoint.

(3) If bds(w) = x y, then both x and y are cycles, and $alph(x \cap y) = alph(x) \cap alph(y)$, where $x \cap y$ is the intersection of x and y.

Therefore comparing two tokens can be done on the basis of their structure, and does not require the use of “best alignment” algorithms.

# Simple syntactic tokens

If one simple word v of length n is in a set S of syntactic tokens, then all simple words of the same length, and no others, are in S. Also no structured words that contain a simple sub-word of length n are in S.

Therefore the only bordered words in S are words w of length $|w| \leq n+1$.

It is easy to check that for any $n \geq 2$,

The union of the set of all simple words v, where $|v| = n$, and the set of all bounded words w, where $|w| \leq n+1$, is a set of syntactic tokens.

Example, A = {a, b, c, d}, n = 3

All tokens are constructed by permutations of letters in the following four words:

| | |
|---|---|
| abc | 24 simple words |
| aaa | 4 structured words |
| aabb | 6 |
| abab | 6 |
| | total = 40 |

The complexity of an algorithm constructing descriptions des(w) for $w \in A^*$ is

O(n) for memory and O(n*|w|) for time,

that is independent of the size of alphabet |A|.

Thus if one wanted to look at the structure of written English texts using bordered and simple tokens, a reasonable choice of n would be the length of a selected long word from a standard English dictionary having all different letters.

For example, choosing the word “unforgivable” would give n = 12.

# Bordered syntactic tokens over a 20-letter alphabet

Introduction

Primary structures of proteins are sequences of amino acids that can be viewed as words in a 20-letter alphabet. So different sets of proteins can be viewed as different languages written in the same alphabet. Looking at sequences of syntactic tokens may be a tool for comparing different sets of proteins.

The number N of syntactic tokens of different lengths L:

| L | N | L | N | L | N |
|---|---|---|---|---|---|
| 3 | 20 | 4 | 1,520 | 5 | 61,560 |
| 6 | 1,860,480 | 7 | 46,512,000 | 8 | 1,004,659,200 |
| 9 | 1.91443392 E10 | 10 | 3.250630656 E11 | 11 | 4.936895309 E12 |
| 12 | 6.704425728 E13 | 13 | 8,112355131 E13 | 14 | 8.688935743 E15 |
| 15 | 8.157945226 E16 | 16 | 6,662899911 E17 | 17 | 4.561691265 E18 |
| 18 | 2.595095475 E19 | 19 | 1.171847801 E20 | 20 | 3.941301253 E20 |
| 21 | 8.78277652 E20 | 22 | 9.731608033 E20 | | |

Tokens of the same length L have the same probability to occur in a randomly chosen word $w \in A^*$.

The expected frequencies F of occurrences of tokens of length L in a randomly constructed word $w \in A^*$ of length $|w| = L$.

| L | F | L | F | L | F |
|---|---|---|---|---|---|
| 3 | .25% | 4 | .95% | 5 | 1.92% |
| 6 | 2.91% | 7 | 3.63% | 8 | 3.92% |
| 9 | 3.74% | 10 | 3.17% | 11 | 2.41% |
| 12 | 1.64% | 13 | .99% | 14 | .53% |
| 15 | .25% | 16 | .1% | 17 | .03% |
| 18 | .01% | 19 | .002% | 20 | 3.8 E-4% |
| 21 | 4.2 E-5% | 22 | 2.3 E-6% | | |

Sum of these frequencies is 26.47%

# Interpretation of tables

If we have a randomly created sequence of w of length 400, we expect the following number of occurrences of approximately 100 tokens of different lengths:

| 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 8 | 12 | 15 | 16 | 15 | 13 | 10 | 7 | 4 | 2 | 1 |

And we would expect that no token occurs twice.

We would also expect that in a sample of 1000 such words, only tokens up to length 7 would occur more than once.

For example, in such a sample we expect to see 16,000 tokens of length 8. But they come from the population of 1,004,659,200 tokens of this length.

We may expect that in a sample taken from one source, the number of repeats among middle lengths tokens will be large. And we may expect that the number of middle-sized tokens common to samples taken from two sources may be used as a measure of their “syntactic similarity”.

But such a measure can be completely meaningless, unless a detailed analysis of the data explains the cause of the observed similarity.

# Final Remarks

1. The idea of syntactic tokens is not new. Similar concepts were discussed in the early nineteen sixties, before Chomsky's generative grammars dominated structural linguistics. (Generative grammars required that minimal syntactic tokens do not overlap)
2. The current definition of syntactic tokens is tentative. And if this concept happens to be useful, it needs to be reconsidered.
3. It would be nice to have a full characterization of all sets of syntactic tokens for a few of the smallest alphabets.

Slides from this talk are available at https://web.nmsu.edu/~pbaggett/notes/notes.html under the title Syntactic Tokens.

Thank you