Cerebellum: Character-based Language Model

Vít Baisa

Natural Language Processing Centre
Faculty of Informatics
Masaryk University

Karlova studánka, RASLAN, 2014

Partially supported by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.

Introduction

The goal of the CBLM is:

- get rid of tokenization and any other rule-based processing
- limit language units not by length but by frequency
- do not work with word units, use bytes as they are universal for all languages represented in computers
- keep language modelling simple
- ...and simpler

Related work

- character-based models not common
- usually used as auxiliary models
- when Markov rule (*n*th order) is used, character-based models lack sufficient context width
- the principle of CBLM proposed in the past: suffix-array language model (SALM, 2007), but on word-level

Suffix tree example

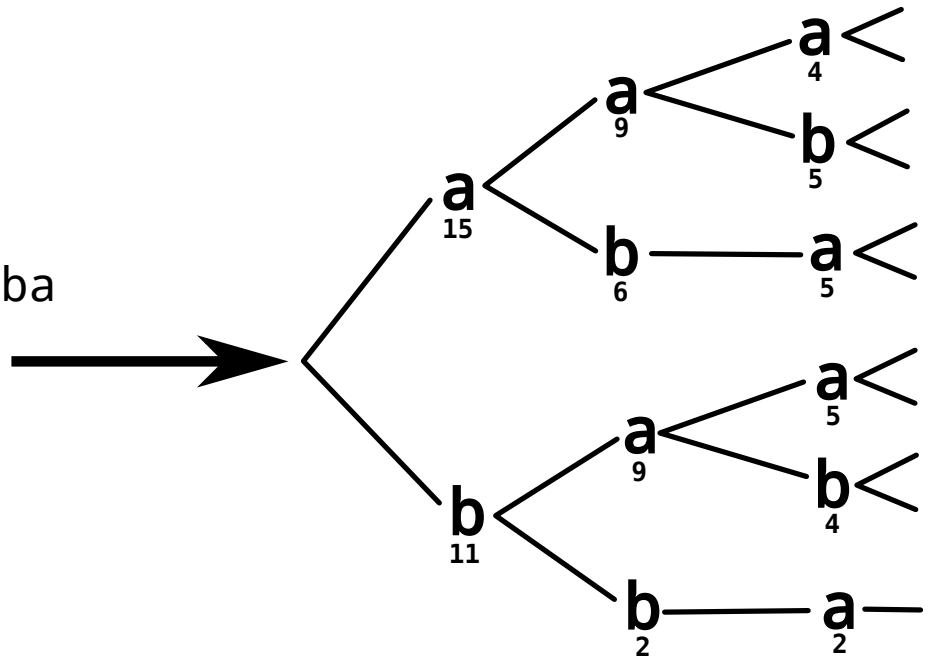
Input is any plain text data, one sentence per line.

I	suffix	SA	sorted suffix	LCP
0	popocatepetl	5	atepetl	0
1	opocatepetl	4	catepetl	0
2	pocatepetl	7	epetl	0
3	ocatepetl	9	etl	1
4	catepetl	11	l	0
5	atepetl	3	ocatepetl	0
6	tepetl	1	opocatepetl	1
7	epetl	8	petl	0
8	petl	2	pocatepetl	1
9	etl	0	popocatepetl	2
10	tl	6	tepetl	0
11	l	10	tl	1

LCP up to 255 (longer sequences not stored).

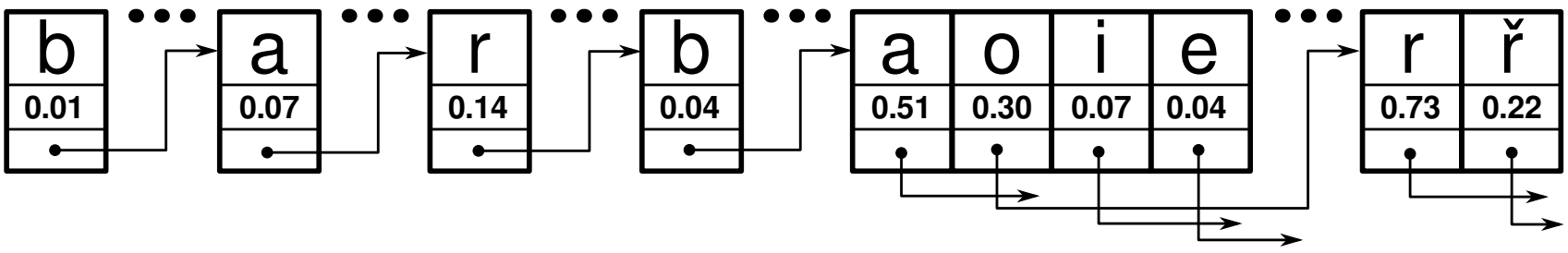
From SA to trie

aaaabaabab	ababbbab
aaaabababa	abbbbbaab
aaabbaaaba	baaabaa
aaabbabab	baaababab
aababba	baaabbba
aababbbab	baabaabbabba
aabbaabab	baabbba
aabbabbab	babaabba
aabbbbba	bababbaaab
abaababbaba	babbbababa
abaabbaba	babbbbaba
abaabbbbab	bbabaaaa
ababbaba	bbabbbab



Parameter N: all sequences occurring $> N\times$ are put to trie.

Trie as stored on disk



Example of a Czech model, prefix *barb*

Language model

Assign probability to any byte sequence:

```
P = ROOT            // pointer to a node in trie
index = 0           // position in input sequence
probability = 1.0
current_path = [] // path from ROOT to P
while index < length(input):
    find input[index] among children(P)
    if not found:
        shorten from left and translate the current_path to ROOT
        until it can be prolonged with input[index] byte
                    // current_path may be emptied
        P = current_path[-1] or ROOT
    else:
        P = found children position
    probability *= probability of P
    current_path.append(P)
    index++
return probability
```

Entropy & perplexity

$$H(m) = -\frac{1}{N}\sum_{i=1}^{N} \log p(b_i)$$

$$PP(m) = 2^{H(m)}$$

Model_N	Test	Size	H	PP
BNC_2	Alice	330 M	2.286	4.879
BNC_3	Alice	212 M	2.294	4.904
BNC_2	1984	330 M	**1.664**	3.170
BNC_3	1984	212 M	1.671	3.184
SYN2000_5	1984	362 M	1.837	3.574
csWiki_3	1984	300 M	1.850	3.607
csWeb_4	1984	312 M	**1.571**	2.972

Example random sentences

English First there is the fact that he was listening to the sound of the shot and killed in the end a precise answer to the control of the common ancestor of the modern city of katherine street, and when the final result may be the structure of conservative politics; and they were standing in the corner of the room.

Czech Pornoherečka Sharon Stone se nachází v blízkosti lesa. ¶ Máme malý byt, tak jsem tu zase. ¶ Změna je život a tak by nás nevolili. ¶ Petrovi se to začalo projevovat na veřejnosti. ¶ Vojáci byli po zásluze odměněni pohledem na tvorbu mléka. ¶ Graf znázorňuje utrpení Kristovo, jež mělo splňovat následující kritéria.

Hungarian Az egyesület székhelye: 100 m-es uszonyos gyorsúszásban a következő években is részt vettek a díjat az égre nézve szójaszármazékot. ¶ Az oldal az első lépés a tengeri akvarisztikával foglalkozó szakemberek számára is ideális szállás költsége a vevőt terhelik.

Future work

- design the *learning* process: using a fixed length window, put new data into the model (trie) on-the-fly
- compute relations with gaps (...atka $\rightarrow$ ...ala)
- use these relations to infer segmentation and grammar rules (be able to parse *Žmoulačka chřpíná u trpného hzunáka.*
- mimic short term memory by decaying contribution of nodes visited in the past
- CBLM for translation (...ed $\rightarrow$ ...aný)
- measure entropy and perplexity on standard datasets
- more extrinsic evaluation (NP segmentation, speech-to-text)