07 – Language modelling

IA161 Advanced Techniques of Natural Language Processing

V. Baisa

NLP Centre, FI MU, Brno

October 30, 2017

Language models—what are they good for?

- assigning scores to sequencies of words
- predicting words
- generating text

$\Rightarrow$

- statistical machine translation
- automatic speech recognition
- optical character recognition

Predicting words

Do you speak ...
Would you be so ...
Statistical machine ...
Faculty of Informatics, Masaryk ...
WWII has ended in ...
In the town where I was ...
Lord of the ...

Generating text

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image


A person riding a motorcycle on a dirt road.


Two dogs play in the grass.


A skateboarder does a trick on a ramp.


A dog is jumping to catch a frisbee.


A group of young people playing a game of frisbee.


Two hockey players are fighting over the puck.


A little girl in a pink hat is blowing bubbles.


A refrigerator filled with lots of food and drinks.

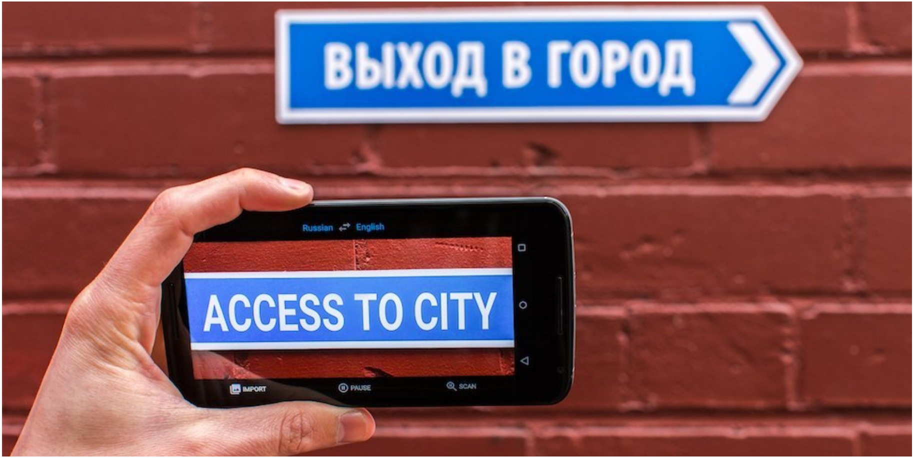

A herd of elephants walking across a dry grass field.


A close up of a cat laying on a couch.


A red motorcycle parked on the side of the road.

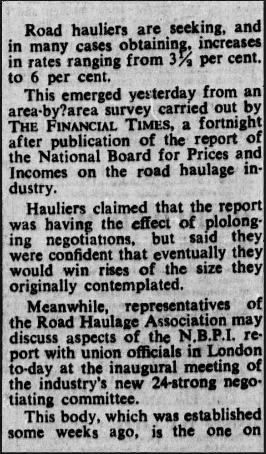

A yellow school bus parked in a parking lot.

MT + OCR


ВЫХОД В ГОРОД
Russian English
ACCESS TO CITY
IMPORT
PAUSE
SCAN

OCR

Raw Article Scan



Road hauliers are seeking, and in many cases obtaining, increases in rates ranging from 3½ per cent. to 6 per cent.

This emerged yesterday from an area-by?area survey carried out by THE FINANCIAL TIMES, a fortnight after publication of the report of the National Board for Prices and Incomes on the road haulage industry.

Hauliers claimed that the report was having the effect of plolonging negotiations, but said they were confident that eventually they would win rises of the size they originally contemplated.

Meanwhile, representatives of the Road Haulage Association may discuss aspects of the N.B.P.I. report with union officials in London to-day at the inaugural meeting of the industry's new 24-strong negotiating committee.

This body, which was established some weeks ago, is the one on

Financial Times 13th July 1965, Page One

3rd Party OCR

ROaR""^'*
i
1.
H
24Wy"
m..
m
.!;
*4
-,ta~
a
·*P
sr;i
r·~·
i
.l~s
24
a·
·

Tesseract OCR
(with default settings)

Road hauliers are seeking, and
in many cases obtunin.
increasesin rates ranging from 3A per cent.
to 6 per cent.
This emerged yesterday from an
area~b ?area survey carried out by
Tue mmcuu. Tums, a fortnight
after publication of the report of
the National Board for Prices and
incomes on the med haulage in-
dustry.
Heuliers claimed that the report
was having the e\ufb01ect of plolong-
ing negotiations. but said they
were con\ufb01dent that eventually they
would win rise: of the tile they
originally contemplated.
Meanwhile. representative: of
the Road Haulage Associntion may
disease aspects of the N.B.P.I. re-
port wrth union of\ufb01ciele in London
to-dey at the inaugural meettng of \u2018
thenndustry\u2018s. new 244trong nego-
: tinting committee. ,
This body. which we! established
some weeks ago. in the one on

Language models – probability of a sentence

- LM is a probability distribution over all possible word sequences.
- What is the probability of utterance of s?

Probability of sentence

p_{LM}(Catalonia President urges protests)
p_{LM}(President Catalonia urges protests)
p_{LM}(urges Catalonia protests President)
…

Ideally, the probability should strongly correlate with fluency and intelligibility of a word sequence.

N-gram models

- an approximation of long sequencies using short n-grams
- a straightforward implementation
- an intuitive approach
- good local fluency

Randomly generated text

“Jsi nebylo vidět vteřin přestal po schodech se dal do deníku a položili se táhl ji viděl na konci místnosti 101,” řekl důstojník.

Hungarian

A társaság kötelezettségeiért kapta a középkori temploma az volt, hogy a felhasználók az adottságai, a felhasználó azonosítása az egyesület alapszabályát.

N-gram models, naïve approach

$$W = w_1, w_2, \cdots, w_n$$

$$p(W) = \prod_i p(w_i | w_1 \cdots w_{i-1})$$

Markov's assumption

$$p(W) = \prod_i p(w_i | w_{i-2}, w_{i-1})$$

$$p(\textit{this is a sentence}) = p(\textit{this}) \times p(\textit{is}|\textit{this}) \times p(a|\textit{this}, \textit{is}) \times p(\textit{sentence}|\textit{is}, a)$$

$$p(a|\textit{this}, \textit{is}) = \frac{|\textit{this is a}|}{|\textit{this is}|}$$

Sparse data problem.

Computing, LM probabilities estimation

Trigram model uses 2 preceding words for probability learning. Using **maximum-likelihood estimation**:

$$p(w_3|w_1,w_2) = \frac{count(w_1,w_2,w_3)}{\sum_w count(w_1,w_2,w)}$$

quadrigram: *(lord, of, the, ?)* ()

w	**count**	$p(w)$
rings	30,156	0.425
flies	2,977	0.042
well	1,536	0.021
manor	907	0.012
dance	767	0.010
...		

Large LM – n-gram counts

How many unique n-grams in a corpus?

order	unique	singletons
unigram	86,700	33,447 (38.6%)
bigram	1,948,935	1,132,844 (58.1%)
trigram	8,092,798	6,022,286 (74.4%)
4-gram	15,303,847	13,081,621 (85.5%)
5-gram	19,882,175	18,324,577 (92.2%)

Corpus: Europarl, 30 M tokens.

Language models smoothing

The problem: an n-gram is missing in the data but is in a *sentence* $\rightarrow$ $p(sentence) = 0$.

We need to assign non-zero p for *unseen data*. This must hold:

$$\forall w . p(w) > 0$$

The issue is more pronounced for higher-order models.

Smoothing: an attempt to amend real counts of n-grams to expected counts in any (unseen) data.

Add-one smoothing (Laplace)

Maximum likelihood estimation assigns p based on

$$p = \frac{c}{n}$$

Add-one smoothing uses

$$p = \frac{c + 1}{n + v}$$

where v is amount of all possible n-grams. That is quite inaccurate since all permutations might outnumber real (possible) n-grams by several magnitudes.

Europarl has 139,000 unique words = 19 G possible bigrams. But it has only 53 M tokens, so maximally 53 M bigrams.

This smoothing overvalues unseen n-grams.

Add-α smoothing

We won't add 1, but α. This can be estimated for the smoothing to be the most just and balanced.

$$p = \frac{c + \alpha}{n + \alpha v}$$

α can be obtained experimentally: we can try several different values and find the best one.

Usually it is very small (0.0001).

Deleted estimation

We can find unseen n-grams in another corpus. N-grams contained in one of them and not in the other help us to estimate general amount of unseen n-grams.

E.g. bigrams not occurring in a training corpus but present in the other corpus million times (given the amount of all possible bigrams equals 7.5 billions) will occur approx.

$$\frac{10^6}{7.5 \times 10^9} = 0.00013\times$$

Good–Turing smoothing

We use *frequency of frequencies*: number of various n-grams which occurr $n\times$.

We use frequency of hapax legomena (singletons in data) to estimate unseen data.

$$r^* = (r+1)\frac{N_{r+1}}{N_r}$$

Especially for n-grams not in our corpus we have

$$r_0^* = (0+1)\frac{N_1}{N_0} = 0.00015$$

where $N_1 = 1.1 \times 10^6$ a $N_0 = 7.5 \times 10^9$ (Europarl).

Example of Good–Turing smoothing (Europarl)

r	**FF**	r^*
0	7,514,941,065	0.00015
1	1,132,844	0.46539
2	263,611	1.40679
3	123,615	2.38767
4	73,788	3.33753
5	49,254	4.36967
6	35,869	5.32929
8	21,693	7.43798
10	14,880	9.31304
20	4,546	19.54487

Smoothing Good-Turing smoothing

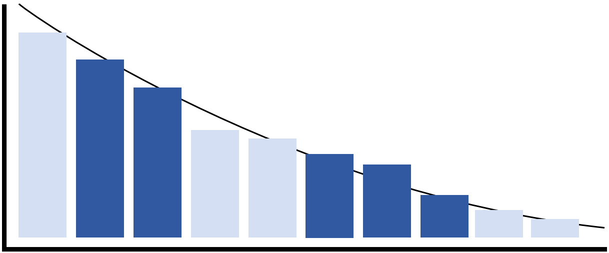


Figure: Missing ranks might be interpolated using standard techniques.

Interpolation and back-off

Previous methods treated all unseen n-grams the same. Consider trigrams

beautiful young girl
beautiful young granny

Despite we don't have any of these in our training data, the former trigram should be more probable.

We will use probability of lower order models, for which we have necessary data:

young girl
young granny
beautiful young

Interpolation

$$p_I(w_3|w_1w_2) = \lambda_1 p(w_3) \times \lambda_2 p(w_3|w_2) \times \lambda_3 p(w_3|w_1w_2)$$

If we have enough data we can trust higher order models more and assign a higher significance to corresponding n-grams.

p_I is probability distribution, thus this must hold:

$$\forall \lambda_n : 0 \leq \lambda_n \leq 1$$
$$\sum_n \lambda_n = 1$$

Quality and comparison of LMs

We need to compare quality of various LM (various orders, various data, smoothing techniques etc.)

1) extrinsic (WER, MT, ASR, OCR) and 2) intrinsic (perplexity) evaluation

A good LM should assign a higher probability to a good (looking) text than to an incorrect text. For a fixed test text we can compare various LMs.

Cross-entropy

$$
\begin{aligned}
H(p_{LM}) &= -\frac{1}{n}\log p_{LM}(w_1, w_2, \ldots w_n) \\
&= -\frac{1}{n}\sum_{i=1}^{n}\log p_{LM}(w_i|w_1, \ldots w_{i-1})
\end{aligned}
$$

Cross-entropy is average value of negative logarithms of words probabilities in testing text. It corresponds to a measure of uncertainty of a probability distribution. **The lower the better**.

A good LM should reach entropy close to real entropy of language. That can't be measured directly but quite reliable estimates exist, e.g. Shannon's game. For English, entropy is estimated to approx. 1.3 bit per letter.

Perplexity

$$PP = 2^{H(p_{LM})}$$

Perplexity is a simple transformation of cross-entropy.

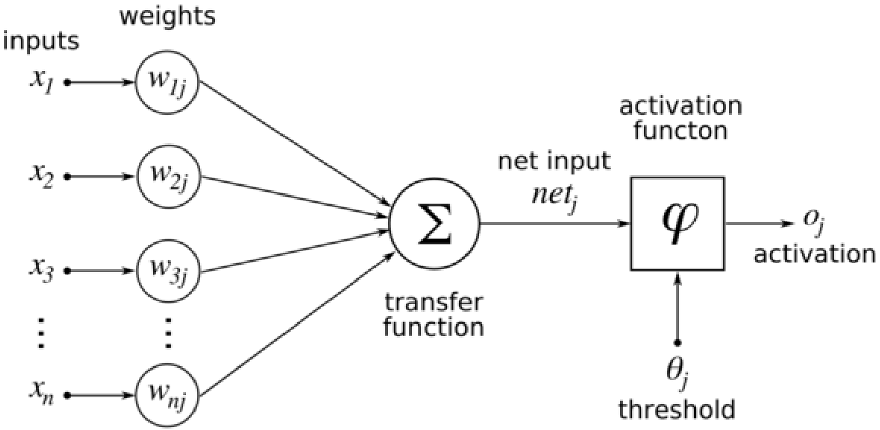
A good LM should not waste p for improbable phenomena.

The lower entropy, the better $\rightarrow$ the lower perplexity, the better.

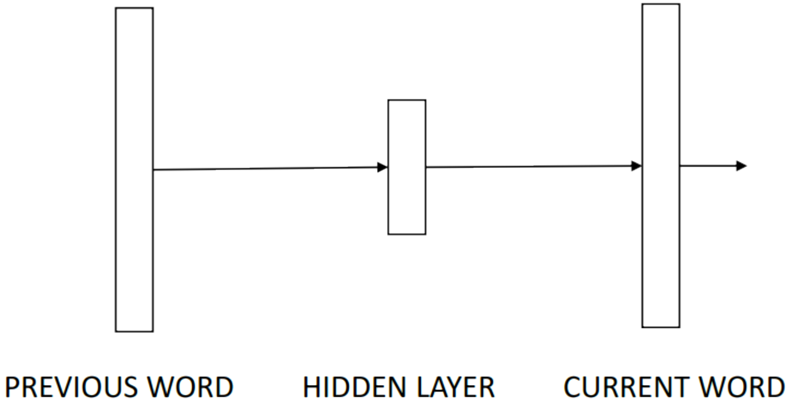
Comparing smoothing methods (Europarl)

method	**perplexity**
add-one	382.2
add-α	113.2
deleted est.	113.4
Good–Turing	112.9

Neuron in artificial neural network



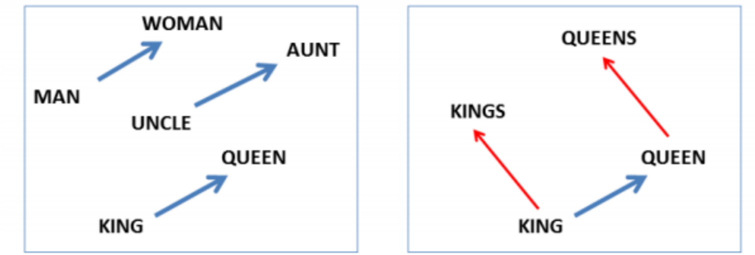
Basic NN



Bigram neural language model.
One-hot representation of words: [0 0 0 0 0 0 1 0 0 0 0]

Distributional Representation of Words

- goal: more compact representation of vectors
- limited dimensionality (500–1000)
- [Mikolov et al., 2013b]
- word vectors capture many linguistic properties (gender, tense, plurality, even semantic concepts like "capital city of")



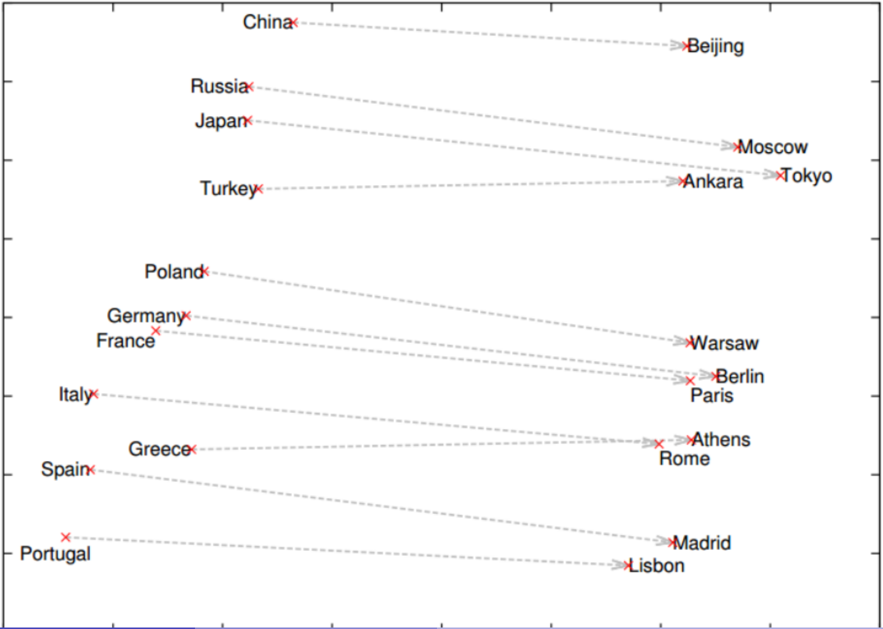
Features: nearest neighbors

	Redmond	Havel	graffiti	capitulate
Collobert NNLM	conyers lubbock keene	plauen dzerzhinsky osterreich	cheesecake gossip dioramas	abdicate accede rearm
Turian NNLM	McCarthy Alston Cousins	Jewell Arzu Ovitz	gunfire emotion impunity	- - -
Mnih NNLM	Podhurst Harlang Agarwal	Pontiff Pinochet Rodionov	anaesthetics monkeys Jews	Mavericks planning hesitated
Skip-gram (phrases)	Redmond Wash. Redmond Washington Microsoft	Vaclav Havel president Vaclav Havel Velvet Revolution	spray paint grafitti taggers	capitulation capitulated capitulating

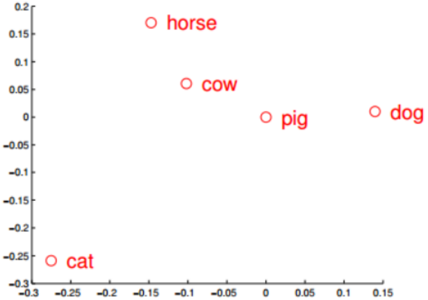
Features: vector arithmetics I

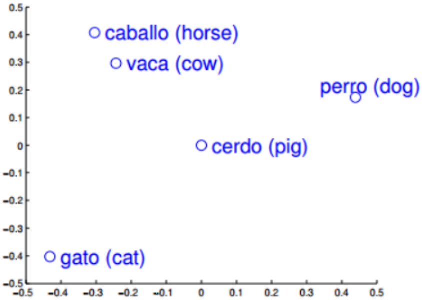
Expression	*Nearest token*
Paris - France + Italy	Rome
bigger - big + cold	colder
sushi - Japan + Germany	bratwurst
Cu - copper + gold	Au
Windows - Microsoft + Google	Android
Montreal Canadiens - Montreal + Toronto	Toronto Maple Leafs

Features: vector arithmetics II



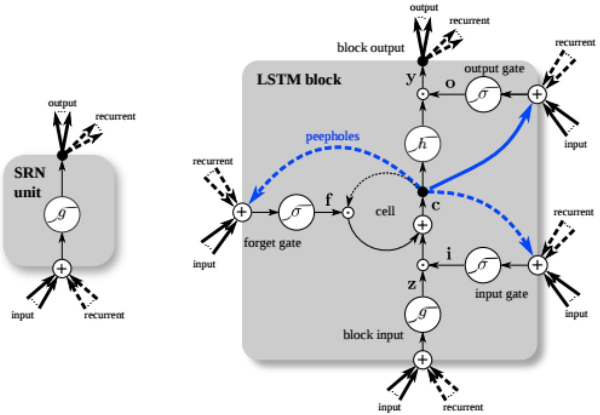
Features: vector arithmetics III





State-of-the-art neural models

The combine advantages of word embeddings and high performance of GPU (vector operations). Terms to remember: recurrent and long-short term memory models.



Best models

Model	Num. Params [billions]	Training Time [hours]	[CPUs]	Perplexity
Interpolated KN 5-gram, 1.1B n-grams (KN)	1.76	3	100	67.6
Katz 5-gram, 1.1B n-grams	1.74	2	100	79.9
Stupid Backoff 5-gram (SBO)	1.13	0.4	200	87.9
Interpolated KN 5-gram, 15M n-grams	0.03	3	100	243.2
Katz 5-gram, 15M n-grams	0.03	2	100	127.5
Binary MaxEnt 5-gram (n-gram features)	1.13	1	5000	115.4
Binary MaxEnt 5-gram (n-gram + skip-1 features)	1.8	1.25	5000	107.1
Hierarchical Softmax MaxEnt 4-gram (HME)	6	3	1	101.3
Recurrent NN-256 + MaxEnt 9-gram	20	60	24	58.3
Recurrent NN-512 + MaxEnt 9-gram	20	120	24	54.5
Recurrent NN-1024 + MaxEnt 9-gram	20	240	24	51.3

References I

Bahl, L. R., Brown, P. F., De Souza, P. V., and Mercer, R. L. (1989).
A tree-based statistical language model for natural language speech recognition.
Acoustics, Speech and Signal Processing, IEEE Transactions on, 37(7):1001–1008.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003).
A neural probabilistic language model.
The Journal of Machine Learning Research, 3:1137–1155.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. (2013).
One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling.
ArXiv e-prints.

References II

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. (2013).
One billion word benchmark for measuring progress in statistical language modeling.
arXiv preprint arXiv:1312.3005.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013).
Scalable modified kneser-ney language model estimation.
In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Kuhn, R. and De Mori, R. (1990).
A cache-based natural language model for speech recognition.
Pattern Analysis and Machine Intelligence, IEEE Transactions on, 12(6):570–583.

References III

Mikolov, T. (2012).
Statistical language models based on neural networks.
Presentation at Google, Mountain View, 2nd April.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a).
Efficient estimation of word representations in vector space.
arXiv preprint arXiv:1301.3781.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b).
Linguistic regularities in continuous space word representations.
In *HLT-NAACL*, pages 746–751.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014).
Show and Tell: A Neural Image Caption Generator.
ArXiv e-prints.

Character-based Language Model

The goal of the CBLM is:

- limit language units not by length but by frequency
- vari-grams
- use bytes as basic units

Suffix tree example

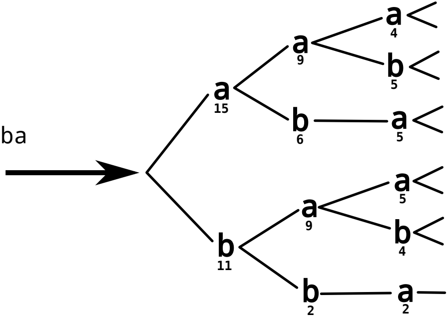
Input is any plain text data, one sentence per line.

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



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

Trie as stored on disk

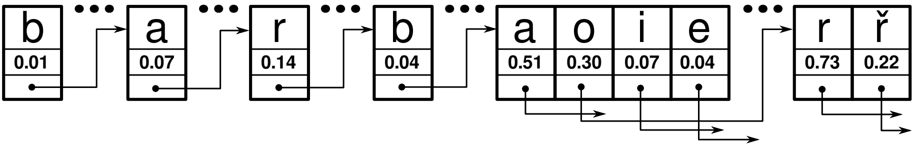


Figure: Example of a Czech model, prefix *barb*

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.