Affisix — Tool for Prefix Recognition

Jaroslava Hlaváčová
hlavacova@ufal.mff.cuni.cz

Michal Hrušecký
michal.hrusecky@seznam.cz

Affisix is a tool for automatic recognition of affixes.
Input: an extensive list of words in a language
Output: segments — candidates for affixes

- prefixes (in the present shape)
- suffixes (another possible use)

Program features

- Support of several methods
- User definable combinations of methods
- Opensource
- Free to use
- Portability to other platforms
- Easy to extend

Properties of prefixes

We have a word:

underestimate

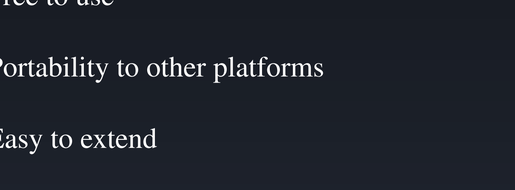
We are searching for border between prefix and the rest of the word.

under|**estimate**

Prefix features:

- The prefix occurs at the beginning of many words.
  Examples: mini|sport, mini|golf, mini|bar
- The prefix is possible to replace with another prefix within the language.
  Examples: mini|shop × maxi|shop × euro|shop × bio|shop

Method of Squares



Square is a quadruple of word segments $\langle p_1, p_2, s_1, s_2 \rangle$ such that $p_1 :: s_1 \in \Omega$, $p_2 :: s_1 \in \Omega$, $p_1 :: s_2 \in \Omega$ and $p_2 :: s_2 \in \Omega$, where Ω stands for the set of all words in a language and :: stands for concatenation.

Many squares with a segment p ⇒
⇒higher chance that p is a prefix

Normalization

Number of squares can be very high, that is why we use its logarithm as a normalized value.

Entropy methods

$$H(p) = -\sum_{s_i \in S} p(s_i|p) \log_2 p(s_i|p)$$

- S is a set of possible end segments,
- $p(s_i|p)$ is the probability that s_i is an end segment, given the beginning segment p

Measure of uncertainty in what would follow. If the entropy is high, the string p is quite common segment which can continue with many other segments to form a word.

Forward Entropy

From left to right.
In forward entropy, we start with the first letter and then prolong the beginning segment of the word.

Backward Entropy

From right to left.
Pretty similar to previous method, but we are starting at the end of the word and proceeding to its beginning.

Difference Entropy

Difference between two subsequent entropies (forward or backward).
In difference entropy method, instead of deciding purely on the entropy value, we use entropy growth. There is naturally great uncertainty in the beginning (because we still have many words), so we try to take different scale in the beginning and at the end.

Normalization

Normalized value = the value divided by the maximal value obtained from the method.

Combination of methods

In our case — sum of normalized values of forward entropy, forward difference entropy and number of squares.
There is number of other possible combinations, for instance using only two best methods, additional weighting.

Results

Data: Czech National Corpus SYN2000.
Methods:

fentr forward entropy,

dfentr difference forward entropy,

fsqrs squares method,

combi combination of normalized values of all methods.

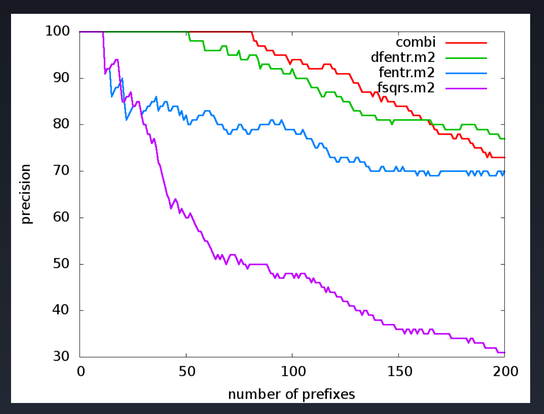
Suffix **m2** means additional condition, that the prefix has to be shorter than half of the word.

Precision

Precision of the methods for 10/50/150/200 best prefix candidates:

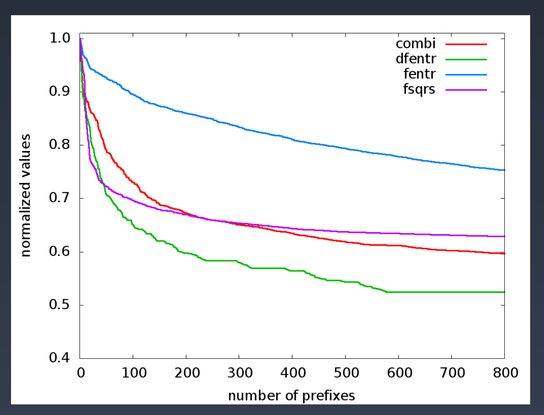
method	10	50	100	150	200
combi	100%	100%	94%	84%	73%
dfentr	100%	100%	84%	71%	60%
dfentr.m2	100%	100%	92%	81%	77%
fentr	100%	82%	79%	70%	70%
fentr.m2	100%	82%	79%	70%	70%
fsqrs	100%	60%	48%	36%	31%
fsqrs.m2	100%	60%	48%	36%	31%

Relationship between number of prefix candidates (horizontal axis) and the precision (vertical axis):



Comparison of normalized results

In this graph, there is shown how fast results from different methods drop.



Summary

In our test, the best way to discover prefixes was followed by difference entropy method. However, there are many ways how to combine these methods and other combinations can possibly bring better results. Thus, this is in our future plans together with experiments on other languages.

Additional Informations

For testing the methods and experimenting on your data in your language, download affisix from

`http://affisix.sf.net`

If you encounter any problem or you just want to contact us, join our mailing list at sourceforge

`affisix-users@lists.sourceforge.net`

The first 85 results sorted according to the combination of the methods:

	prefix	combi	||fentr||	||dfentr||	||fsqrs||
1.	super	2.682	0.967	0.978	0.737
2.	pseudo	2.602	0.963	0.961	0.677
3.	mikro	2.532	0.921	0.917	0.693
4.	sebe	2.505	0.898	0.849	0.757
5.	rádoby	2.480	0.983	1.000	0.496
6.	deseti	2.437	0.917	0.865	0.653
7.	mimo	2.423	0.968	0.801	0.653
8.	hyper	2.420	0.906	0.894	0.619
9.	anti	2.393	0.942	0.706	0.744
10.	roz	2.390	0.922	0.530	0.937
11.	severo	2.384	0.966	0.839	0.578
12.	jiho	2.370	0.944	0.848	0.577
13.	makro	2.366	0.935	0.886	0.544
14.	elektro	2.364	0.931	0.797	0.635
15.	jedno	2.361	0.923	0.726	0.711
16.	nízko	2.359	0.940	0.919	0.499
17.	mega	2.341	0.904	0.861	0.575
18.	vnitro	2.332	0.939	0.862	0.530
19.	spolu	2.327	0.921	0.711	0.695
20.	dvou	2.317	0.929	0.659	0.728
21.	euro	2.311	0.918	0.791	0.601
22.	velko	2.309	0.857	0.820	0.631
23.	auto	2.306	0.896	0.737	0.673
24.	ultra	2.306	0.873	0.888	0.544
25.	devíti	2.304	0.858	0.833	0.612
26.	pře	2.302	0.816	0.531	0.954
27.	dvoj	2.298	0.935	0.665	0.697
28.	čtyř	2.288	0.802	0.776	0.709
29.	mezi	2.283	0.933	0.683	0.667
30.	multi	2.281	0.857	0.809	0.614
31.	proti	2.270	0.964	0.538	0.766
32.	foto	2.269	0.889	0.738	0.642
33.	pěti	2.267	0.921	0.651	0.693
34.	mnoho	2.250	0.921	0.701	0.627
35.	krypto	2.227	0.871	0.794	0.561
36.	video	2.223	0.870	0.755	0.596
37.	vše	2.219	0.936	0.648	0.634
38.	několika	2.214	0.863	0.714	0.636
39.	kvazi	2.207	0.854	0.850	0.502
40.	východo	2.194	0.936	0.722	0.536
41.	dvaceti	2.183	0.871	0.688	0.623
42.	šéf	2.182	0.889	0.687	0.605
43.	západo	2.178	0.929	0.723	0.524
44.	třista	2.171	0.821	0.835	0.513
45.	padesáti	2.162	0.861	0.708	0.592
46.	čtvrt	2.146	0.876	0.678	0.591
47.	vysoko	2.145	0.891	0.756	0.498
48.	čtyřiceti	2.129	0.850	0.689	0.589
49.	staro	2.128	0.948	0.567	0.612
50.	imuno	2.124	0.833	0.769	0.521
51.	trans	2.112	0.837	0.662	0.611
52.	středo	2.110	0.927	0.594	0.588
53.	novo	2.108	0.941	0.571	0.595
54.	samo	2.107	0.913	0.543	0.650
55.	tele	2.106	0.888	0.600	0.617
56.	neuro	2.104	0.874	0.662	0.568
57.	čtver	2.099	0.796	0.756	0.546
58.	patnácti	2.097	0.825	0.675	0.596
59.	porno	2.095	0.848	0.769	0.478
60.	hydro	2.083	0.819	0.685	0.578
61.	mini	2.081	0.904	0.510	0.666
62.	mnoha	2.074	0.867	0.646	0.560
63.	rychlo	2.069	0.874	0.663	0.531
64.	infra	2.069	0.841	0.803	0.424
65.	modro	2.061	0.904	0.649	0.507
66.	šesti	2.058	0.904	0.477	0.676
67.	dvěstě	2.056	0.758	0.771	0.526
68.	termo	2.056	0.909	0.603	0.543
69.	inter	2.056	0.862	0.548	0.644
70.	šestnácti	2.053	0.823	0.651	0.577
71.	jedenácti	2.040	0.841	0.622	0.576
72.	dolno	2.040	0.860	0.794	0.385
73.	cyklo	2.039	0.867	0.613	0.558
74.	stereo	2.034	0.789	0.780	0.464
75.	psycho	2.034	0.845	0.595	0.593
76.	cyber	2.032	0.828	0.841	0.362
77.	pětadvaceti	2.030	0.782	0.690	0.557
78.	celo	2.024	0.931	0.470	0.621
79.	dvanácti	2.023	0.820	0.587	0.615
80.	pětiset	2.022	0.800	0.636	0.585
81.	cyto	2.021	0.797	0.728	0.495
82.	hotov	2.020	0.738	0.674	0.608
83.	černo	2.008	0.872	0.599	0.536
84.	bičov	2.001	0.693	0.704	0.603
85.	transport	2.000	0.691	0.703	0.605