# Finding the Best Name for a Set of Words Automatically

Pavel Rychlý

Natural Language Processing Centre Faculty of Informatics, Masaryk University Botanická 68a, 602 00 Brno, Czech Republic pary@fi.muni.cz

**Abstract.** Many natural language processing applicatons use clustering or other statistical methods to create sets of words. Such sets group together words with similar meaning and in many cases humans can find an appropriate term quickly. On the other hand computers represent such sets with a meaningless number or ID. This paper proposes an algorithm for automatic finding of names of word sets. It provides result examples as a simple evaluation of the method.

Keywords: names of word sets, naming clusters, distributional thesaurus

### 1 Introduction

There are many applications in natural language processing which process words or lemmas and create some sets of words. Usually it is done via some type of clustering but they could be done using many different statistical methods.

As an example of such applications see Figure 1, it presents an thesaurus for word *milk* in the Sketch Engine system [1]. Thesaurus is computed automatically using a distributional similarity method [2]. The individual words which are similar to the given word (*milk*) are clustered using a bottom up clustering. The front words of each cluster is the word with the highest similarity score in the cluster.

The Sketch Engine thesaurus is based on the Word Sketches. These are one page collocational behavior of a word, an exampel of a Word Sketch for verb *break* is displayed in Figure 2. In is used mainly in lexicography and language learning. A Word Sketch provides lists of collocations devided into several grammatical relations. On the Figure 2, some collocations are clustered using the same technique as in the Thesarus.

The final example is from LDA-frames project [3], Figure 3. LDA-frames is an unsupervised approach to identifying semantic frames from semantically unlabelled text corpora. There are many frame formalisms but most of them suffer from the problem that all frames must be created manually and the set of semantic roles must be predefined. The LDA-Frames approach, based on the

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2014, pp. 77–81, 2014. © NLP Consulting 2014

<b>British National Corpus freq =</b> <u>4692</u> (41.8 per million)					
Lemma	Score	Freq	Cluster		
<u>meat</u>	0.227	3690	<u>fruit</u> [0.177, 4989] <u>vegetable</u> [0.164, 2714] <u>potato</u> [0.16, 2458] <u>bean</u> [0.134, 1744] <u>rice</u> [0.126, 1537] <u>tomato</u> [0.114, 1465]		
<u>coffee</u>	0.222	6372	wine [0.221, 7123] <u>tea</u> [0.202, 8256] <u>beer</u> [0.199, 3629] <u>drink</u> [0.19, 6655]		
<u>juice</u>	0.207	1883	<u>salt</u> [0.128, 3263]		
<u>cream</u>	0.201	3221	bread [0.198, 3668] sugar [0.196, 3685] cheese [0.195, 2918] butter [0.19, 2062] chocolate [0.153, 2316]		
egg	0.191	6071			
<u>oil</u>	0.173	10126	<u>coal</u> [0.108, 5302] gas [0.101, 8082]		
food	0.171	20774	fish [0.134, 10322] goods [0.11, 10052] product [0.106, 21606]		
<u>soup</u>	0.17	1405	sauce [0.137, 1597] salad [0.112, 1394]		
water	0.144	34246	<u>blood</u> [0.133, 9780]		
<u>cake</u>	0.143	3666	biscuit [0.13, 1567] sandwich [0.109, 1769]		
stuff	0.137	6629	<u>meal</u> [0.114, 6532]		

Fig. 1: Thesaurus of *milk* in the Sketch Engine

Latent Dirichlet Allocation, avoids both these problems by employing statistics on a syntactically tagged corpus. The only information that should be given is a number of semantic frames and a number of semantic roles to be identified.

From all these examples we can see that many clusters clearly define one common meaning. A native speaker could easily choose a single word name for such cluster. This paper presents an algorithm to find such name automatically.

# 2 Proposed Method

The proposed method exploits the distributional thesaurus data which provide a list of similar words for a given word. The algorithm works as follows:

- 1. for each word in the given set find a list of top similar words in the thesaurus
- 2. sum the score for each of similar words across all given words
- 3. add 1 to the sums for each input words (the most similar word for any word is the word itself)
- 4. sort similar words according to the sums of scores
- 5. display the top items from the list

### 3 Evaluation

To our knowledge, there are no evaluation data available. We are going to prepare such gold data as a future work. As a simple form of evaluation we list results of the algorithm on our test data. They are presented in Table 1.

mill

<u>243</u> <u>105</u>	<b>3.6</b> 9.12	subject	<u>5542</u>	5.1						
<u>105</u>		Think		5.I	<u>and/or</u>	<u>377</u>	0.1	pp into-p	<u>872</u>	16.5
		Thief	<u>35</u>	7.63	bend	<u>9</u>	6.11	trot	<u>17</u>	8.84
10	8.42	thief	<u>41</u>	7.46	damage	<u>6</u>	4.93	grin <u>20</u>	<u>58</u>	7.84
10	impasse <u>16</u> stalemate <u>10</u>			7.35	enter	<u>18</u>	4.88	smile <u>38</u>		
<u>499</u>	8.15	fighting	<u>39</u>	7.25	fall <u>18</u>	<u>35</u>	4.27	gallop	<u>6</u>	7.31
arm <u>81</u> finger <u>24</u> neck <u>149</u>			<u>244</u>	7.22	try <u>17</u>			applause	<u>8</u>	7.27
<u>80</u>	7.98	strike 14			make <u>72</u>	<u>80</u>	2.68	run	<u>25</u>	6.2
<u>122</u>	7.87	burglar <u>27</u>	<u>33</u>	7.12	go <u>8</u>			garage	<u>8</u>	6.03
		intruder <u>6</u>						laughter	<u>6</u>	5.62
<u>177</u>	7.67	marriage	<u>72</u>	7.0				song	<u>12</u>	5.05
<u>982</u>	7.61	storm	<u>36</u>	6.98	down	<u>704</u>	8.27	thought 22	<u>28</u>	5.03
agreement <u>34</u> code <u>36</u> contract <u>89</u> pattern 25 record <u>186</u> regulation 21			<u>38</u>	6.96	ир	<u>569</u>	6.81	speech 6		
regulatio	<u></u>	wave	<u>50</u>	6.7	off	<u>146</u>	6.71	flat	<u>11</u>	5.01
<u>52</u>	7.6	fight	<u>34</u>	6.7	in	<u>24</u>	3.9	piece	<u>22</u>	4.79
<u>170</u>	7.46	fire	<u>74</u>	6.53	out	<u>60</u>	3.78	tear	<u>6</u>	4.78
<u>81</u>	7.45	raider <u>17</u>	<u>23</u>	6.44	over	<u>10</u>	3.3	house 76	<u>196</u>	4.63
wrist <u>30</u>										
<u>67</u>	7.4	scuffle	<u>14</u>	6.32	•					
<u>59</u>	7.26	scandal	<u>20</u>	6.24	down			time	12	0.84
<u>186</u>	7.24	blaze	<u>15</u>	6.22	through	<u>193</u>		L		
		row	<u>35</u>	6.15	off	<u>532</u>	8.49			
	499 149 122 177 982 contrac regulatio 52 170 81 67 59	499       8.15         149       80         80       7.98         122       7.87         177       7.67         982       7.61         contract 89         regulation 21         52       7.6         170       7.46         81       7.45         67       7.4         59       7.26	499       8.15       fighting         499       8.15       fighting         149       war 230         80       7.98       strike 14         122       7.87       burglar 27         intruder 6       intruder 6         177       7.67       marriage         982       7.61       storm         contract 89       hell       wave         52       7.6       fight         170       7.46       fire         81       7.45       raider 17         attacker 6       59       7.26         59       7.26       scandal         186       7.24       blaze	$\begin{array}{c c c c c c c c c c c c c c c c c c c $	499       8.15       fighting       39       7.25         149       itighting       39       7.25         80       7.98       strike 14       122         122       7.87       burglar 27       33       7.12         intruder 6       intruder 6       1177       7.67       marriage       72       7.0         982       7.61       storm       36       6.98         contract 89       hell       38       6.96         wave       50       6.7         52       7.6       fight       34       6.7         170       7.46       fire       74       6.53         81       7.45       raider 17       23       6.44         attacker 6       59       7.26       scuffle       14       6.32         59       7.26       scandal       20       6.24       186       7.24       blaze       15       6.22	499       8.15       fighting       39       7.25       fall <u>18</u> 149       war <u>230</u> 244       7.22       try <u>17</u> 80       7.98       strike <u>14</u> make <u>72</u> 122       7.87       burglar <u>27</u> 33       7.12       go <u>8</u> 177       7.67       marriage       72       7.0       part trans         982       7.61       storm       36       6.98       down         982       7.61       storm       36       6.98       down         contract <u>89</u> hell       38       6.96       up         regulation <u>21</u> wave       50       6.7       off         170       7.46       fire       74       6.53       out         170       7.46       scuffle       14       6.32       part intrans         67       7.4       scuffle       14       6.32       down         186       7.24       blaze       15       6.22       through off	499       8.15       fighting       39       7.25       fall $18$ 35         149       ifighting       39       7.25       fall $18$ 35         149       war $230$ $244$ 7.22       try $17$ make $72$ 80         122       7.87       burglar $27$ 33       7.12       go 8       make $72$ 80         177       7.67       marriage       72       7.0       go 8       make $72$ 80         982       7.61       storm       36       6.98       up       569         60       vave       50       6.7       6.7       off       146         52       7.6       fight       34       6.7       out       60         81       7.45       raider $17$ 23       6.44       out       60         81       7.26       scandal       20       6.24       part intrans       4343         659       7.26       scandal       20       6.24       fown       159         186       7.24       blaze       15       6.22       off       159	4998.15fighting $39$ 7.25fall $18$ $35$ $4.27$ $149$ war $230$ $244$ 7.22fall $18$ $35$ $4.27$ $80$ 7.98strike $14$ make $72$ $80$ $2.68$ $122$ 7.87burglar $27$ $33$ $7.12$ make $72$ $80$ $122$ 7.87burglar $27$ $33$ $7.12$ $go 8$ $177$ 7.67marriage $72$ $7.0$ $part trans15209827.61storm366.98down7048.27go garwave506.7off1466.71contract 89hell386.96up5696.81wave506.7off1466.71527.6fight346.7out603.781707.46fire746.53over103.3817.45scuffle146.32partntrans434322.4677.4scuffle146.32partntrans434322.4677.26scandal206.24down15919.391867.24blaze156.22eff5228.92$	Image: second secon	499 $8.15$ $100$ $100$ $100$ $110$ <

Fig. 2: Word sketch of verb *break* in the Sketch Engine

	1
input word set	output top names
oil coal gas	fuel-n 0.696
	energy-n 0.536
Britain Scotland Europe England	country-n 4.189
	area-n 3.308
apple pear orange	fruit-n 2.145
	thing-n 1.441
procedure study analysis method programme	system-n 5.367
	work-n 4.959
pint bottle litre gallon	glass-n 2.371
	water-n 2.258
meat fruit vegetable potato	food-n 3.291
	fish-n 2.803
village town	city-n 0.611
	area-n 0.478

Table 1: Result of the algorithm on test data.

#### Pavel Rychlý

EAT

	SUB	JECT	OBJECT			
	22	22	40			
0.554086 frame 1166	0.794216 0.010335 0.007963 0.005797 0.004342 0.003409 0.002687 0.002519 0.002307 0.002215	person people one man who woman child that all someone	0.085888 0.046396 0.01947 0.01947 0.01726 0.016846 0.015189 0.013256 0.012289 0.012151	food meal egg breakfast lunch dinner fish meat potato cake		
	15	152		40		
0.128011 frame 622	0.027104 0.026926 0.023538 0.023181 0.016049 0.014979 0.013374 0.01266 0.011947 0.011769	bird dog animal fish cat child people prey man horse	0.085888 0.046396 0.01947 0.01947 0.01726 0.016846 0.015189 0.013256 0.012289 0.012151	food meal egg breakfast lunch dinner fish meat potato cake		

Fig. 3: Verb eat in LDA-frames

### 4 Interface

The algorithm is implemented as a command line script. It is written in Python and uses the Sketch Engine API to access the thesaurus data. We assume that after more finetuning the algorithm will be included into the Sketch Engine system. An example of a usage is at Figure 4.

```
$ clustname.py bnc2 bnc-hyper n Britain Scotland Europe England
country-n 4.18891489506
area-n 3.50870908797
year-n 3.5038651228
London-n 3.2635447681
world-n 3.13785666227
```

Fig. 4: An example of the clustname.py tool usage.

# 5 Conclusions

We have proposed an algorithm for finding names for a set of words. The implementation is mostly language and corpus independent and works quite well for many test data.

**Acknowledgements** This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013 and by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.

# References

- 1. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. Proceedings of Euralex (2004) 105–116 http://www.sketchengine.co.uk.
- Rychlý, P., Kilgarriff, A.: An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics (2007) 41–44
- Materna, J.: LDA-Frames: An Unsupervised Approach to Generating Semantic Frames. In Gelbukh, A., ed.: Proceedings of the 13th International Conference CICLing 2012, Part I. Volume 7181 of Lecture Notes in Computer Science., Springer Berlin / Heidelberg (2012) 376–387