Longest-commonest match for word sketches

Vít Baisa, †‡ Adam Kilgarriff, †Pavel Rychlý, †‡ Miloš Jakubíček †‡

Lexical Computing, † Masaryk University ‡ Brighton, UK & Brno, CZ

July 23, 2015 CL 2015, Lancaster

Outline

- 1 Sketch Engine
- 2 Word sketches & Longest-commonest match
- Conclusions

Sketch Engine

- corpus management system
- web service (including API)
- platform for providing language resources
- widely used for
 - lexicography purposes
 - Harper Collins, Oxford University Press, Cambridge University Press, Macmillan, . . .
 - linguistic and language technology teaching and research at universities
 - more than 100 academic institutions worldwide
 - dozens of thousands of individuals
 - language modelling (IT/LT companies)

Sketch Engine features

- **concordancing**, sorting, sampling, wordlists, collocation lists
- full regular-expression searching
- support for parallel corpora, virtual sub- and supercorpora
- handles billion-word (80 G+) corpora smoothly
- word sketches: one-page summaries of a word's grammatical and collocational behaviour
- distributional thesaurus
- keywords extraction, terms extraction
- Corpus Architect: user corpora
 - uploaded by users
 - created by WebBootCaT

Sketch Engine languages

By June 2015 more than **400 corpora** for **82 languages**:

- 100+ corpora having more than 100 million tokens
- 30+ corpora having more than 1 billion tokens
 - In 2010 a series of TenTen (10¹⁰) corpora started
- 60+ languages with a PoS-tagged corpus
- 42 languages with word sketches
- 26 languages with integrated tagger for tagging user corpora
- parallel corpora: EUROPARL, DGT, OPUS, ...

Users

Lexicographers

Sketch Engine

- Researchers
- Teachers
- Language Learners
- Translators
- Terminologists
- Copywriters

Learning how words are used

reading concordances

Reading concordances

Query resource 13,141 (117.14 per million)

1	bncdoc#1	room is now a well established and valuable	resource	for undergraduates in particular. Susan
2	bncdoc#1	at Brown University. The college's	resources	are as much if not more in its members
3	bncdoc#1	Somerville to make maximum use of its extra	resources	. Somerville Lawyers The College Librarian
4	bncdoc#1	mini pupillages. The new group provides a valuable new	resource	for those turning to law later in their
5	bncdoc#1	financial trouble. Please help. With additional	resources	the Fund could do more. The Secretary of
6	bncdoc#1	projects, and begin a survey of the space and	resources	which should be provided if new purpose built
7	bncdoc#1	because trust funds usually direct their	resources	to new short term projects rather than
8	bncdoc#2	expertise which exists already in the Corine	resource	, a project which ends this year, is in
9	bncdoc#2	Forests " Trained amateurs" to tap rainforest	resources	American biologists have called for the
10	bncdoc#2	populations to help tap potential medical	resources	of tropical forests. Thomas Eisner
11	bncdoc#8	The College has the ideal environment and	resources	to run major conferences and seminars.
12	bncdoc#8	international context (managing markets,	resources	, people, information) The enhancement of
13	bncdoc#8	exchanges and joint research.	RESOURCES	THE LIBRARY The Library forms part of
14	bncdoc#8	corporate benefit of information systems and	resources	'. College staff have regular contact with
15	bncdoc#8	and implementation of organisational human	resource	policies in both the public and private
16	bncdoc#10	geographers are able to use the library	resources	of the Bodleian, other departments of the
17	bncdoc#10	distribution of human settlements; natural	resource	use and technological innovation; forms
18	bncdoc#10	inform students of the specialist educational	resources	available in the different departments
19	bncdoc#10	Coin Room, offers a particularly important	resource	for both research and teaching in Classical
20	bncdoc#10	(a) annual estimates; (b) allocation of	resources	and accommodation; (c) junior academic

Learning how words are used

- reading concordances
 - slow and tedious

Learning how words are used

- reading concordances
 - slow and tedious
- reviewing collocations / word sketches
 - quick overview "one-page summary of..."

Word sketch

resource (noun) British National Corpus freq = 12658 (112.8 per million)

modifier	6477	1.5	object of	3285	2.2	modifies	1906	0.5	subject of	512	0.6
scarce	<u>163</u>	9.53	allocate	<u>194</u>	9.58	allocation	<u>135</u>	9.42	devote	28	7.69
natural	<u>321</u>	8.94	pool	<u>39</u>	8.43	implication	<u>46</u>	7.09	consume	4	5.36
limited	<u>187</u>	8.86	exploit	<u>64</u>	8.23	management	<u>153</u>	6.98	tie	<u>6</u>	4.87
financial	249	8.3	divert	<u>38</u>	7.86	defense	<u>7</u>	6.68	last	<u>4</u>	4.6
mineral	89	8.19	deploy	<u>31</u>	7.67	Stonier	<u>6</u>	6.65	back	<u>5</u>	4.5
additional	<u>107</u>	7.92	devote	<u>44</u>	7.64	utilisation	<u>7</u>	6.63	stretch	4	4.29
valuable	<u>74</u>	7.86	concentrate	<u>62</u>	7.35	committee	<u>132</u>	6.49	result	<u>6</u>	3.93
extra	88	7.53	utilise	22	7.28	centre	<u>158</u>	6.4	depend	<u>6</u>	3.84
human	<u>134</u>	7.38	conserve	<u>17</u>	7.09	allocator	<u>5</u>	6.4	limit	<u>5</u>	3.59
renewable	<u>33</u>	7.31	lack	<u>37</u>	7.0	depletion	<u>6</u>	6.21	match	<u>3</u>	3.58
adequate	<u>49</u>	7.28	reallocate	<u>13</u>	6.98	pack	<u>17</u>	6.2	share	<u>6</u>	3.55
non-renewable	<u>25</u>	6.97	mobilise	<u>13</u>	6.83	investigator	<u>8</u>	6.17	earn	<u>3</u>	3.55
existing	<u>53</u>	6.68	mobilize	<u>13</u>	6.79	column	20	6.16	enable	<u>7</u>	3.54
finite	<u>22</u>	6.66	distribute	<u>29</u>	6.73	constraint	<u>14</u>	6.14	remain	<u>12</u>	3.5

- easier than reading concordances
 - resource modifiers of r. scarce
 - resource verbs with r. as object allocate
 - ...easy bits

- easier than reading concordances
 - resource modifiers of r. scarce
 - resource verbs with r. as object allocate
 - ... easy bits
- ... but not always as informative

- easier than reading concordances
 - resource modifiers of r. scarce
 - resource verbs with r. as object allocate
 - ...easy bits
- ... but not always as informative
 - put objects of p. phone (BNC)

- easier than reading concordances
 - resource modifiers of r. scarce
 - resource verbs with r. as object allocate
 - ... easy bits
- ... but not always as informative
 - put objects of p. phone (BNC)
 - ⇒ put the phone down

- easier than reading concordances
 - resource modifiers of r. scarce
 - resource verbs with r. as object allocate
 - ...easy bits
- ... but not always as informative
 - put objects of p. phone (BNC)
 - ⇒ put the phone down
 - resource nouns modified by r. department (SkELL)

- easier than reading concordances
 - resource modifiers of r. scarce
 - resource verbs with r. as object allocate
 - ... easy bits
- ... but not always as informative
 - put objects of p. phone (BNC)
 - ⇒ put the phone down
 - resource nouns modified by r. department (SkELL)
 - ⇒ human resources department

- easier than reading concordances
 - resource modifiers of r. scarce
 - resource verbs with r. as object allocate
 - ... easy bits
- ... but not always as informative
 - put objects of p. phone (BNC)
 - ⇒ put the phone down
 - resource nouns modified by r. department (SkELL)
 - ⇒ human resources department
 - resource words and/or r. conservation (SkELL)

- easier than reading concordances
 - resource modifiers of r. scarce
 - resource verbs with r. as object allocate
 - ... easy bits
- ... but not always as informative
 - put objects of p. phone (BNC)
 - ⇒ put the phone down
 - resource nouns modified by r. department (SkELL)
 - ⇒ human resources department
 - resource words and/or r. conservation (SkELL)
 - ⇒ state department of conservation and natural resources

- sometimes it is necessary to see concordance examples to understand the word sketch
- ⇒ slows down the workflow

Understanding word sketches

- sometimes it is necessary to see concordance examples to understand the word sketch
- ⇒ slows down the workflow

make it easier by providing most common concordance examples for the headword-relation-collocation triple

Understanding word sketches

- sometimes it is necessary to see concordance examples to understand the word sketch
- ⇒ slows down the workflow

- make it easier by providing most common concordance examples for the headword-relation-collocation triple
- "commonest match"

Understanding word sketches

- sometimes it is necessary to see concordance examples to understand the word sketch
- ⇒ slows down the workflow

- make it easier by providing most common concordance examples for the headword-relation-collocation triple
- "commonest match"
- . . . the longest available ⇒ longest-commonest match (LCM)

Longest-commonest match (LCM)

- headword relation collocation
- resource noun modifier department
- ⇒ human resources department

LCM algorithm

- get all concordance matches starting with headword and ending with collocate (or vice versa if applies)
- does any match occur in at least 1/4 of the concordance lines?
 - yes it is an LCM candidate, try extending the phrase to the left or right and repeat with this subset of concordances
 - no quit
- simple, fast, language independent

LCM algorithm

- get all concordance matches starting with headword and ending with collocate (or vice versa if applies)
- does any match occur in at least 1/4 of the concordance lines?
 - yes it is an LCM candidate, try extending the phrase to the left or right and repeat with this subset of concordances
 - no quit
- simple, fast, language independent
- ... but no fast enough for online computation ⇒ LCM preindexed as part of word sketches
 - triples: (headword, relation, collocation)
 - 5-tuples: + (headword position, collocation position)
 - n-tuples: + (longest commonest match words)

LCM integration in Sketch Engine

put (verb) British National Corpus + Commonest Match freq = 67,367

•				
down	part_trans	2,534	9.94	put down
forward	modifier	1,720	11.56	put forward
up	part_intrans	<u>1,176</u>	7.83	to put up with
just	modifier	<u>832</u>	8.77	just put
in	part_intrans	<u>796</u>	8.88	put in
pressure	o bjec t	<u>519</u>	8.20	put pressure on
then	modifier	<u>438</u>	8.05	and then put
head	object	<u>387</u>	6.92	put his head
thing	object	<u>382</u>	6.57	put things
end	object	<u>327</u>	6.80	to put an end to
off	part_intrans	320	7.76	be put off
place	pp_in-p	<u>241</u>	6.61	put in place
right	np_adj_comp	<u>217</u>	7.96	put it right
phone	object	213	7.47	put the phone down
hand	part_out-a_obj	<u>167</u>	5.92	put out a hand
lot	object	<u>147</u>	6.30	put a lot of
use	pp_to-p	<u>145</u>	6.33	put to good use
kettle	object	<u>142</u>	7.15	put the kettle on

LCM examples

- baby in BNC
- baby in SkELL
- test in SkELL
- resource in SkELL

LCM evaluation

- not a classification task
- some LCM is clearly better than no LCM
- what should the threshold be?
 - two annotators
 - options 1/2, 1/4, 1/6
 - 1/6 the best, people prefer longer matches
- primitive filtering of LCM (e.g. punctuation) easy to do but might not be desirable

Conclusions

- understanding collocations can be made easier by providing the most common concordance examples
- examples = Longest-commonest match
- simple algorithm
- integrated into Sketch Engine, anywhere where a word sketch is (soon)