Evaluation of the Sketch Engine Thesaurus on Analogy Queries

Pavel Rychlý

Natural Language Processing Centre
Faculty of Informatics, Masaryk University

December 3, 2016

Contents

Sketch Engine Thesaurus

Lemma	Score	Freq
king	0.242	16,899
prince	0.213	6,355
charles	0.189	8,952
elizabeth	0.177	3,567
edward	0.176	6,484
mary	0.173	6,870
gentleman	0.171	6,274
lady	0.170	11,905
husband	0.167	11,669
sister	0.167	8,062
mother	0.164	27,536
princess	0.160	2,944
father	0.159	23,824
wife	0.157	18,308
brother	0.155	11,049
henry	0.151	6,699
daughter	0.150	11,216
anne	0.149	4,386

queen *(noun)*
British National Corpus (BNC) freq = 7,872 (70.10 per million)

Thesaurus evaluation

Gold standard

Source	Most similar words to *queen*
serelex	king, brooklyn, bowie, prime minister, mary, bronx, rolling stone, elton john, royal family, princess
Thesaurs.com	monarch, ruler, consort, empress, regent, female ruler, female sovereign, queen consort, queen dowager
SkE on BNC	king, prince, charles, elizabeth, edward, mary, gentleman, lady, husband, sister, mother, princess, father
SkE on enTenTen08	princess, prince, king, emperor, monarch, lord, lady, sister, lover, ruler, goddess, hero, mistress, warrior
word2vec on BNC	princess, prince, Princess, king, Diana, Queen, duke, palace, Buckingham, duchess, lady-in-waiting, Prince
powerthesaurus.org	empress, sovereign, monarch, ruler, czarina, queen consort, king, queen regnant, princess, rani, queen regent

Analogy queries

- evaluation of word embeddings (word2vec)
- "a is to a^* as b is to b^*", where b^* is hidden

Analogy queries

- evaluation of word embeddings (word2vec)
- "a is to a^* as b is to b^*", where b^* is hidden
- syntactic: *good* is to *best* as *smart* is to *smarter*
- semantic: *Paris* is to *France* as *Tokyo* is to *Japan*
- very high agreement by humans:

Analogy queries

- evaluation of word embeddings (word2vec)
- "a is to a^* as b is to b^*", where b^* is hidden
- syntactic: *good* is to *best* as *smart* is to *smarter*
- semantic: *Paris* is to *France* as *Tokyo* is to *Japan*
- very high agreement by humans:
  Berlin –

Analogy queries

- evaluation of word embeddings (word2vec)
- "a is to a^* as b is to b^*", where b^* is hidden
- syntactic: *good* is to *best* as *smart* is to *smarter*
- semantic: *Paris* is to *France* as *Tokyo* is to *Japan*
- very high agreement by humans:
  Berlin – Germany

Analogy queries

- evaluation of word embeddings (word2vec)
- "a is to a^* as b is to b^*", where b^* is hidden
- syntactic: *good* is to *best* as *smart* is to *smarter*
- semantic: *Paris* is to *France* as *Tokyo* is to *Japan*
- very high agreement by humans:
  Berlin – Germany
  London –

Analogy queries

- evaluation of word embeddings (word2vec)
- "a is to a^* as b is to b^*", where b^* is hidden
- syntactic: *good* is to *best* as *smart* is to *smarter*
- semantic: *Paris* is to *France* as *Tokyo* is to *Japan*
- very high agreement by humans:
  Berlin – Germany
  London – England / Britain / UK ?

Analogy queries

- evaluation of word embeddings (word2vec)
- "a is to a^* as b is to b^*", where b^* is hidden
- syntactic: *good* is to *best* as *smart* is to *smarter*
- semantic: *Paris* is to *France* as *Tokyo* is to *Japan*
- very high agreement by humans:
  Berlin – Germany
  London – England / Britain / UK ?
- best match for linear combination of vectors:
  $\arg\max_{b^* \in V} cos(b^*, a^* - a + b)$

From vectors to similarity

- $cos(x, y) = \frac{v_x \cdot v_y}{\sqrt{v_x \cdot v_x}\sqrt{v_y \cdot v_y}}$
- $\arg\max_{b^* \in V} cos(b^*, a^* - a + b) =$

From vectors to similarity

- $cos(x, y) = \frac{v_x \cdot v_y}{\sqrt{v_x \cdot v_x}\sqrt{v_y \cdot v_y}}$
- $\arg\max_{b^* \in V} cos(b^*, a^* - a + b) =$
  $\arg\max_{b^* \in V}(cos(b^*, a^*) - cos(b^*, a) + cos(b^*, b))$
  (CosAdd)

From vectors to similarity

- $cos(x, y) = \frac{v_x \cdot v_y}{\sqrt{v_x \cdot v_x}\sqrt{v_y \cdot v_y}}$
- $\arg\max_{b^* \in V} cos(b^*, a^* - a + b) =$
  $\arg\max_{b^* \in V}(cos(b^*, a^*) - cos(b^*, a) + cos(b^*, b))$
  (CosAdd)
- $\arg\max_{b^* \in V} \frac{cos(b^*,a^*)cos(b^*,b)}{cos(b^*,a)}$
  (CosMul)
- SkE uses Jaccard similarity instead of cosine similarity:
  JacAdd, JacMul

Results

Results on capital-common-countries question set (462 queries)

	BNC		SkELL	
	count	percent	count	percent
CosAdd	58	12.6	183	39.6
CosMul	99	21.4	203	43.9
JacAdd	32	6.9	319	69.0
JacMul	57	12.3	443	95.9
word2vec	159	34.4	366	79.2

Implementation

First implementation in a few lines of bash:

```
join -j 2 <(dumpthes bnc2 woman-n|sort -k2)
     <(dumpthes bnc2 king-n|sort -k2)
|join -2 2 -a 1 - <(dumpthes bnc2 man-n|sort -k2)
|awk '$4==""{$4=0.05};{print $1, $2*$3/($4+0.0001)}'
|sort -t ' ' -k2r |head
```

```
man-n 1.53483
queen-n 0.276912
parent-n 0.229166
father-n 0.227446
mother-n 0.224917
```

Implementation

First implementation in a few lines of bash:

```
join -j 2 <(dumpthes bnc2 woman-n|sort -k2)
     <(dumpthes bnc2 king-n|sort -k2)
|join -2 2 -a 1 - <(dumpthes bnc2 man-n|sort -k2)
|awk '$4==""{$4=0.05};{print $1, $2*$3/($4+0.0001)}'
|sort -t ' ' -k2r |head

man-n 1.53483
queen-n 0.276912
parent-n 0.229166
father-n 0.227446
mother-n 0.224917
```

Full implementation (in Go) is 150× faster than word2vec

Results on other corpora

More English corpora, using JacMul

Corpus	size (M)	correct
BNC	112	57
SkELL	1,520	443
araneum maius (LCL sketches)	1,200	224
enclueweb16	16,398	448
ententen 08	3,268	0
ententen 12	12,968	0
ententen 13	22,878	439