Evaluation of the Sketch Engine Thesaurus on Analogy Queries

Pavel Rychlý

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

Abstract. Recent research on vector representation of words in texts bring new methods of evaluating distributional thesauri. One of such methods is the task of analogy queries. We evaluated the Sketch Engine thesaurus on a subset of analogy queries using several similarity options. We show that Jaccard similarity is better than the cosine one for bigger corpora, it even substantially outperforms the word2vec system.

Key words: distributional thesaurus, analogy queries

1 Introduction

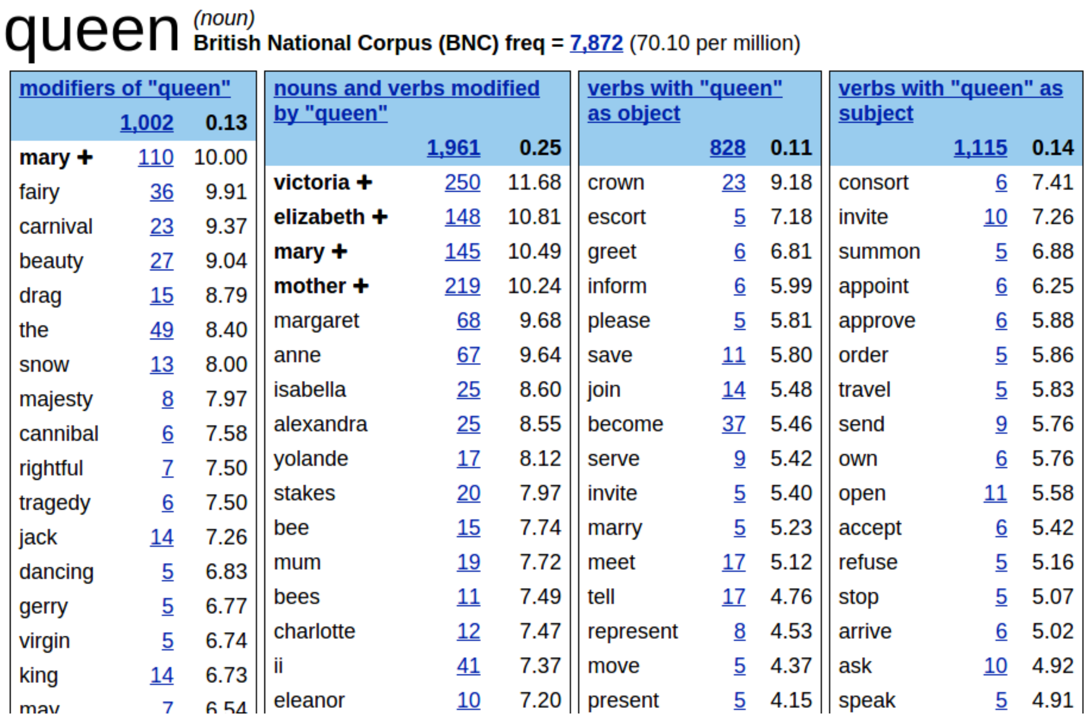
A thesaurus contains words grouped together according to similarity of meaning. Like dictionaries, they are hard and expensive to produce. There is a long history of projects and tools trying to produce a thesaurus automatically from large text corpora. Such tools produce a list of similar words for each given word. Similar usually means words occurring in same or similar contexts. That could include not only synonyms and antonyms (as expected in human-created thesauri) but also words from the same class (like animals) or hypernyms and hyponyms. Such data sets (usually called distributional thesauri) are helpful for humans as another type of dictionary but they also useful in many natural language processing tasks.

There are many different approaches how to build a thesaurus from a text corpus with many parameters and options for each method. To compare which algorithm/settings is better there are methods for evaluating thesauri from the very beginning of building automatic thesauri in 1965 [1]. Thesaurus evaluation is discussed in more details in the next section.

The Sketch Engine (SkE) [2] is a corpus management system with several unique features. One of the most important feature (which also gave the name to the whole system) is a word sketch. It is a one page overview of grammatical and collocational behaviour of a given word. It is an extension of the general collocation concept used in corpus linguistics in that they group collocations according to particular grammatical relation (e.g. subject, object, modifier etc.). An example of word sketch for noun *queen* on British National

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016, pp. 147–152, 2016. © Tribun EU 2016

Corpus (BNC) [3] is in Figure 1. Another features of the Sketch Engine is a thesaurus. It is based on word sketches, similarity of two words is derived from the intersection of collocations in respective grammatical relations of both words. An example of the thesaurus result for noun *queen* on BNC is in Figure 2. More technical details of the Sketch Engine thesaurus computation are in Section 3.



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

modifiers of "queen"	1,002	0.13
mary +	110	10.00
fairy	36	9.91
carnival	23	9.37
beauty	27	9.04
drag	15	8.79
the	49	8.40
snow	13	8.00
majesty	8	7.97
cannibal	6	7.58
rightful	7	7.50
tragedy	6	7.50
jack	14	7.26
dancing	5	6.83
gerry	5	6.77
virgin	5	6.74
king	14	6.73
may	7	6.54

nouns and verbs modified by "queen"	1,961	0.25
victoria +	250	11.68
elizabeth +	148	10.81
mary +	145	10.49
mother +	219	10.24
margaret	68	9.68
anne	67	9.64
isabella	25	8.60
alexandra	25	8.55
yolande	17	8.12
stakes	20	7.97
bee	15	7.74
mum	19	7.72
bees	11	7.49
charlotte	12	7.47
ii	41	7.37
eleanor	10	7.20

verbs with "queen" as object	828	0.11
crown	23	9.18
escort	5	7.18
greet	6	6.81
inform	6	5.99
please	5	5.81
save	11	5.80
join	14	5.48
become	37	5.46
serve	9	5.42
invite	5	5.40
marry	5	5.23
meet	17	5.12
tell	17	4.76
represent	8	4.53
move	5	4.37
present	5	4.15

verbs with "queen" as subject	1,115	0.14
consort	6	7.41
invite	10	7.26
summon	5	6.88
appoint	6	6.25
approve	6	5.88
order	5	5.86
travel	5	5.83
send	9	5.76
own	6	5.76
open	11	5.58
accept	6	5.42
refuse	5	5.16
stop	5	5.07
arrive	6	5.02
ask	10	4.92
speak	5	4.91

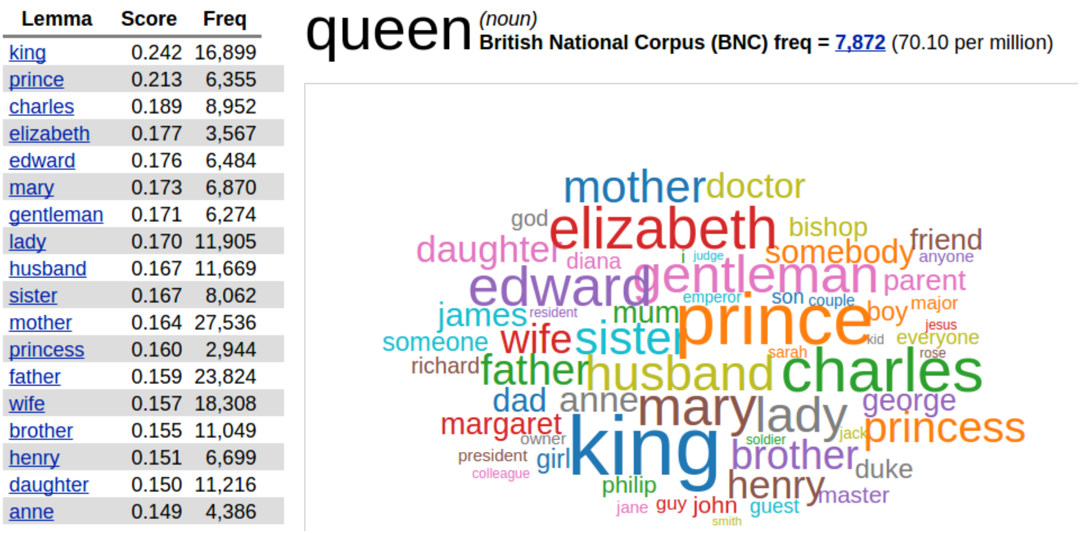
Fig. 1: Word Sketch of word *queen* on British National Corpus.

2 Thesaurus Evaluation

The first methods of evaluating thesaurus quality was based on gold standards – data prepared by several annotators. They contain a list of word pairs together with a numeric or quality assignment of their similarity. There are several problems with such data:

- some gold standards do not distinguish between similarity and relatedness (*money* – *bank*: score 8.5 out of 10 in WordSim353 data set [4])
- some gold standards do not provide any measure of similarity [5]

It is very hard for a human to decide which ordering of similar words is better. As an illustration, the Table 1 lists most similar words for noun *queen* from several sources.



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

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

Fig. 2: Sketch Engine Thesaurus for word *queen* on British National Corpus.

Table 1: Most similar words for noun *queen* from different sources

Source	Most similar words to *queen*
serelex[5]	king, brooklyn, bowie, prime minister, mary, bronx, rolling stone, elton john, royal family, princess, monarch, manhattan, prince, harper, head of state, iron maiden, kiss, paul mccartney, abba, hendrix
Thesaurs.com	monarch, ruler, consort, empress, regent, female ruler, female sovereign, queen consort, queen dowager, queen mother, wife of a king
SkE on BNC	king, prince, charles, elizabeth, edward, mary, gentleman, lady, husband, sister, mother, princess, father, wife, brother, henry, daughter, anne, doctor, james
SkE on enTenTen08	princess, prince, king, emperor, monarch, lord, lady, sister, lover, ruler, goddess, hero, mistress, warrior, knight, priest, chief, god, maiden, brother
word2vec on BNC	princess, prince, Princess, king, Diana, Queen, duke, palace, Buckingham, duchess, lady-in-waiting, Prince, coronation, empress, Elizabeth, hrh, Alianor, Edward, King, bride
powerthesaurus.org	empress, sovereign, monarch, ruler, czarina, queen consort, king, queen regnant, princess, rani, queen regent, female ruler, grand duchess, infanta, kumari, maharani, crown princess, kunwari, shahzadi, malikzadi

With recent research on vector representation of words [6] they came new methods of evaluating such representations. One of such methods is the task of analogy queries. Each query is in form "a is to a^* as b is to b^*", where b^* is hidden and the system mush guess it. There are several types of analogy in the evaluation data set, we can divide them into two classes: morpho-syntactic ("*good* is to *best* as *smart* is to *smarter*") and semantic ("*Paris* is to *France* as *Tokyo* is to *Japan*"). Most of such queries are easy to answer by humans and there is almost 100 % agreement.

Using vector representation of words, a vector of real numbers is assigned to each word. The analogy query is answered by finding the closest word to the result of vector $a^* - a + b$. The distance of vectors is computed as the cosine similarity of two vectors:

$$cos(x,y) = \frac{v_x \cdot v_y}{\sqrt{v_x \cdot v_x}\sqrt{v_y \cdot v_y}}$$

where v_x and v_y are the respective vectors of words x and y. The analogy query is answered by:

$$\arg\max_{b^* \in V} cos(b^*, a^* - a + b)$$

3 Sketch Engine Thesaurus

As we mentioned in the first section the Sketch Engine thesaurus is based on word sketches. It is computed during corpus compilation. For each word, all words with similarity above a threshold are stored together with the similarity score. An efficient algorithm is used to compute the whole $N \times N$ matrix [7].

The question is whether we can use the vector arithmetics for analogy queries together with Sketch Engine thesaurus. At first sight we cannot use it because we have no vectors. But we can derive the vectors from word sketches. Each collocation in a grammatical relation could be one dimension of the vector. The association score of the collocation is the value of that dimension in the vector. But there are two main differences:

- The dimension of word vectors in word2vec system (and also in all others) is only 100 – 1000, the dimension of vectors from word sketches goes up to millions.
- Similarity in SkE thesaurus is computed using Jaccard similarity instead of the cosine one.

Fortunately, vector arithmetic could be interpreted in different way, using CosAdd [8]:

$$\arg\max_{b^* \in V} cos(b^*, a^* - a + b) =$$

$$\arg\max_{b^* \in V} (cos(b^*, a^*) - cos(b^*, a) + cos(b^*, b))$$

It means that we are finding word b^* that is close to a^* and b and far from a. We can also use multiplication instead of addition and use CosMul:

$$\arg\max_{b^* \in V} \frac{cos(b^*, a^*)cos(b^*, b)}{cos(b^*, a)}$$

In our experiments we have defined two other methods with the cosine similarity substituted to the Jaccard one: JacAdd, JacMul.

4 Evaluation

We made the evaluation on two English corpora: BNC and SkELL. BNC (around 100 million tokens) is rather small for semantic relations, SkELL [9] (around 1.5 billion tokens) is large enough. We have selected only one analogy relation (*capital-common-countries*) from the analogy data set because many of the words from other relations have small number of hits in our corpora.

The results are summarised in Table 2. All experiments were evaluated on 462 analogy queries. The table lists number of successful answers and the respective percentage (accuracy) for each configuration.

Table 2: Results on capital-common-countries question set

	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

5 Conclusions

The analogy queries is a very good task for evaluating distributional thesauri. Our results confirm the well-known fact that more texts provides better results. It seems that the cosine similarity gives better results than the Jaccard similarity (SkE default) for smaller corpora. For bigger corpora, the Jaccard similarity is better than the cosine one.

For smaller corpora, word2vec is clearly better option but Sketch Engine thesaurus substantially outperforms word2vec for bigger corpora.

Acknowledgments. This work has been partly supported by the Grant Agency of CR within the project 15-13277S.

References

1. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Communications of the ACM **8**(10) (1965) 627–633
2. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. Proceedings of Euralex (2004) 105–116 `http://www.sketchengine.co.uk`.
3. Aston, G., Burnard, L.: The BNC handbook: Exploring the British National Corpus with SARA. Edinburgh University Press (1998)

4. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: The concept revisited. In: Proceedings of the 10th international conference on World Wide Web, ACM (2001) 406–414
5. Panchenko, A., Morozova, O., Fairon, C., et al.: A semantic similarity measure based on lexico-syntactic patterns. In: Proceedings of KONVENS 2012. (2012)
6. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
7. 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
8. Levy, O., Goldberg, Y., Ramat-Gan, I.: Linguistic regularities in sparse and explicit word representations. In: CoNLL. (2014) 171–180
9. Baisa, V., Suchomel, V.: Skell: Web interface for english language learning. In: Eighth Workshop on Recent Advances in Slavonic Natural Language Processing. (2014) 63–70