Evaluation of Czech Distributional Thesauri

Pavel Rychlý

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

Abstract. Distributional thesauri play a big role in current approaches to natural language applications. There are many ways how to build them and there is no clear way how to compare them which one is better. There are data sets for thesaurus quality evaluation for several languages, the ones for Czech are hard to use for several reasons. This paper proposes a new data set for Czech which is easy to use in different environments.

Keywords: distributional thesaurus, evaluation, outlier detection

1 Introduction

A thesaurus lists words with similar meaning for any given word. There is a long history of algorithms for automatic generation of distributional thesauri based on co-occurrence of words in a very large text. These algorithms find words occurring in same or similar contexts. That could include synonyms, antonyms (as one can expect in a human-created thesaurus) but also words from the same class (like animals) or hypernyms and hyponyms. With the availability of huge text collections, these data sets have good coverage of words and provide important information about words in a language. That is the reason they are successfully used in many natural language applications.

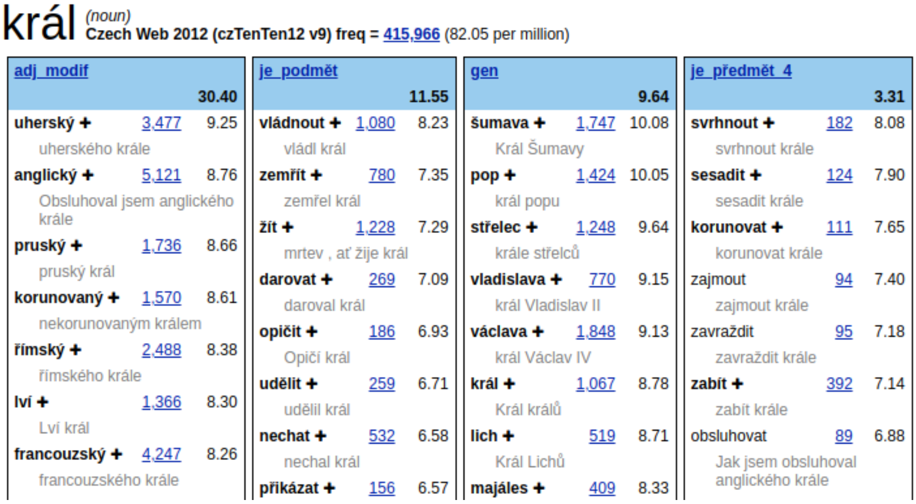
To compare which algorithm and/or settings of building a thesaurus is better there are methods for evaluating thesauri from the very beginning of building automatic thesauri in 1965 [7]. Thesaurus evaluation is discussed in more details in the next section.

The Sketch Engine (SkE) [4] 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 behavior 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.). The system is language independent and this paper will deal with Czech corpora and data generated by the system from them.

An example of word sketch for noun *král (king)* on CzTenTen12 [10] is in Figure 1. The Sketch Engine provides also a thesaurus. It is based on word sketches, similarity of two words is computed as the intersection of collocations

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019, pp. 137–142, 2019. © Tribun EU 2019

in respective grammatical relations of both words. An example of the thesaurus result for noun *král* is in Figure 2. More details about the algorithms behind the thesaurus computation can be find in [9].



král (noun) Czech Web 2012 (czTenTen12 v9) freq = 415,966 (82.05 per million)

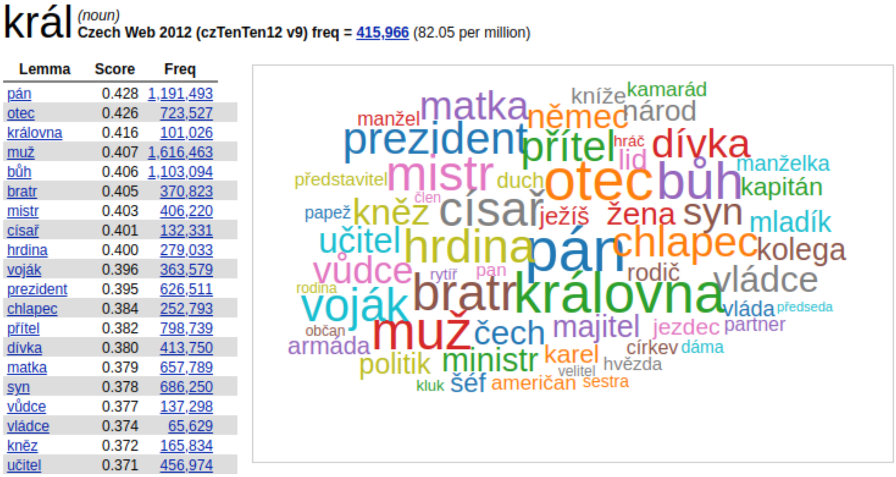
adj_modif		30.40
uherský +	3,477	9.25
uherského krále		
anglický +	5,121	8.76
Obsluhoval jsem anglického krále		
pruský +	1,736	8.66
pruský král		
korunovaný +	1,570	8.61
nekorunovaným králem		
římský +	2,488	8.38
římského krále		
lví +	1,366	8.30
Lví král		
francouzský +	4,247	8.26
francouzského krále		

je_podmět		11.55
vládnout +	1,080	8.23
vládl král		
zemřít +	780	7.35
zemřel král		
žít +	1,228	7.29
mrtev , ať žije král		
darovat +	269	7.09
daroval král		
opičit +	186	6.93
Opičí král		
udělit +	259	6.71
udělil král		
nechat +	532	6.58
nechal král		
přikázat +	156	6.57

gen		9.64
šumava +	1,747	10.08
Král Šumavy		
pop +	1,424	10.05
král popu		
střelec +	1,248	9.64
krále střelců		
vladislava +	770	9.15
král Vladislav II		
václava +	1,848	9.13
král Václav IV		
král +	1,067	8.78
Král králů		
lich +	519	8.71
Král Lichů		
majáles +	409	8.33

je_předmět_4		3.31
svrhnout +	182	8.08
svrhnout krále		
sesadit +	124	7.90
sesadit krále		
korunovat +	111	7.65
korunovat krále		
zajmout	94	7.40
zajmout krále		
zavraždit	95	7.18
zavraždit krále		
zabít +	392	7.14
zabít krále		
obsluhovat	89	6.88
Jak jsem obsluhoval anglického krále		

Fig. 1: Word Sketch of word *král (king)* on CzTenTen12 corpus.



král (noun) Czech Web 2012 (czTenTen12 v9) freq = 415,966 (82.05 per million)

Lemma	Score	Freq
pán	0.428	1,191,493
otec	0.426	723,527
královna	0.416	101,026
muž	0.407	1,616,463
bůh	0.406	1,103,094
bratr	0.405	370,823
mistr	0.403	406,220
císař	0.401	132,331
hrdina	0.400	279,033
voják	0.396	363,579
prezident	0.395	626,511
chlapec	0.384	252,793
přítel	0.382	798,739
dívka	0.380	413,750
matka	0.379	657,789
syn	0.378	686,250
vůdce	0.377	137,298
vládce	0.374	65,629
kněz	0.372	165,834
učitel	0.371	456,974

Fig. 2: Sketch Engine Thesaurus for word *král (king)* on CzTenTen12 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 [3])
- some gold standards do not provide any measure of similarity [6]
- usually, inter annotator agreement within human annotators is low.

Examples of different lists of similar words are discussed in [8].

The high attention to thesauri usage came from recent systems for computing vector representation of words [5], [1]. They bring a new methods of evaluating such representations. The most popular is the task of analogy queries. Each query is in form "*A* is to *B* as *C* is to *D*", where *D* is hidden and the system must guess it. The description of several variants of the task and the application of the task for the Sketch Engine thesaurus is in [8].

The biggest problem of this task is under-specification of a query. In many cases, humans are able to answer a question correctly because they know the query type. For example in the query *"Germany" is to "Berlin" as "France" is to "?"*, the right answer is *"Paris"* because the query is in the set *Capitals*. But Berlin could be the biggest city or the city where the respective president was born or anything else. These queries are also sensitive to the size of the training texts, they contains quite rare name entities. The evaluation therefore focuses on not so important parts of a language.

The most reliable evaluation seems to be so called outlier detection, proposed in [2]. The query in this evaluation set contains a list of words and the task is to find an outlier of that list, the word which is the least "similar" to the other words in the list. The outlier is computed from a thesaurus as a word for which the rest of the given list of words is the most compact. The compactness is defined as the sum (or average) of similarities of all word pairs in a cluster.

3 Evaluating Outlier Detection

The paper introducing this evaluation contains the whole data set. The outlier detection queries are defined in a form of two sets:

- positive set of 8 words forming a well understood cluster,
- negative set of 8 words with outliers.

It is possible to create 8 different queries from each such set pair. One word from the negative set and the whole positive set defines one such query. The article [2] contains 8 pairs of word sets. That mean it is possible to generate $8 \times 8 = 64$ individual queries.

The evaluation of a thesaurus computes two numbers:

- Accuracy – the percentage of successfully answered queries,
- Outlier Position Percentage (OPP) Score – average percentage of the right answer (Outlier Position) in the list of possible clusters ordered by their compactness.

Outlier Position (OP) is a number from 0 to the number of words in the query (9 in the original data set). 0 means the worst guess, the maximum means the right answer. Therefore OPP 100 % indicates all hits which means 100 % accuracy.

OOP provides more fine grained evaluation than accuracy. For incorrect answers it differentiates the position of the right answer in the ordered list.

There are several problems with the original data set for evaluating Czech thesauri. The crucial one is that it contains only English words. That is the main reason we have created our new data set. The second problem is that the evaluation uses the exact word forms in the queries. That is usually not a big problem for English because thesaurus is usually compiled for word forms. Also words in the queries are in the basic form in almost all cases. In Czech there is some times no obvious basic form. For example, many systems use masculine form of adjectives as basing form. But it is not intuitive if the most common collocation is in a different gender. The second problem is that many words are ambiguous, there are several different lemmata, in some cases even in different part of speech.

We have solved this problem by modification of the evaluation script. The original procedure computes the compactness of all possible clusters. In our modification there is a preparation phase for each cluster where we find the best matching lemma (or lemma + part of speech, depending on the thesauri) for each word in the query.

The last problem creates multiword units in the queries. These are handled for thesauri using word vectors as the sum of individual words of the multiword. This technique cannot be used for a Sketch Engine thesaurus. The new feature for computing thesaurus for multiword units is in development, we have skipped the queries containing multiwords in the evaluation of for this paper.

4 The New Data Set

To use outlier detection on Czech corpora we have prepared a new data set of outlier detection lists of Czech words. During the creation of the data, two annotators compiled 48 clusters of words in four different languages. Then they translated each cluster to other languages. The result contains each cluster in five languages: Czech, Slovak, English, German, French. We have found many errors (typos, bad translations) in the data, this paper deals with the Czech part only.

An example of the clusters from the new data set is listed in table 1. It shows two clusters in both Czech and English variant. The first cluster contains words which are not in basing form as adjectives (they use feminine gender) or they are ambiguous (nouns or adjectives). In the second cluster (Electronics) one can see an example of multiword unit (*mp3 player*).

Table 1: Example of two data set clusters – Colors and Electronics

Colors		Electronics	
Czech	English	Czech	English
červená	red	televize	television
modrá	blue	reproduktor	speaker
zelená	green	notebook	laptop
žlutá	yellow	tablet	tablet
fialová	purple	mp3 přehrávač	mp3 player
růžová	pink	mobil	phone
oranžová	orange	rádio	radio
hnědá	brown	playstation	playstation
dřevěná	wooden	blok	notebook
skleněná	glass	sešit	workbook
temná	dark	kniha	book
zářivá	bright	CD	CD
pruhovaný	striped	energie	energy
puntíkovaný	dotted	světlo	light
smutná	sad	papír	paper
nízká	low	ráno	morning

5 Evaluation

We made the evaluation on three Czech corpora: Czes2 (460 million tokens), czTenTen12 (5 billion tokens) [10], csTenTen17 (12 billion tokens) [11].

We have selected only clusters without multiword units, 9 clusters, these form 72 queries. The results are summarized in Table 2. The czTenTen12 corpus was evaluated with Sketch Engine thesaurus and also with word vectors compiled by FastText. We have also included prebuild model from Common Crawl.

Table 2: Evaluation of Czech thesauri on 72 queries of the new data set

	OOP	Accuracy
Czes2	92.2	70.8
czTenTen12	93.4	79.2
csTenTen17	94.3	81.9
czTenTen12 (fasttext)	97.7	87.5
cc.cs	98.1	95.8

6 Conclusions

The outlier detection is probably the best task for evaluating distributional thesauri. This paper describes the new data set of Czech outlier detection lists. We have used this data set on several Czech corpora and we think that this evaluation is suitable for developing new methods and/or optimizing parameters of computing distributional thesauri. We will add more languages to the next version of the new data set.

The comparison of Sketch Engine thesaurus and words vectors generated by FastText shows that FastText provides better results.

Acknowledgements This work has been partly supported by the Czech Science Foundation under the project GA18-23891S.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
2. Camacho-Collados, J., Navigli, R.: Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. pp. 43–50 (2016)
3. 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. pp. 406–414. ACM (2001)
4. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. Proceedings of Euralex pp. 105–116 (2004), `http://www.sketchengine.co.uk`
5. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
6. Panchenko, A., Morozova, O., Fairon, C., et al.: A semantic similarity measure based on lexico-syntactic patterns. In: Proceedings of KONVENS 2012 (2012)
7. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Communications of the ACM **8**(10), 627–633 (1965)
8. Rychlỳ, P.: Evaluation of the sketch engine thesaurus on analogy queries. In: RASLAN. pp. 147–152 (2016)
9. 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. pp. 41–44. Association for Computational Linguistics (2007)
10. Suchomel, V.: Recent czech web corpora. In: RASLAN. pp. 77–83 (2012)
11. Suchomel, V.: cstenten17, a recent czech web corpus. In: RASLAN. pp. 111–123 (2018)