# A Distributional Multi-word Thesaurus in Sketch Engine

Miloš Jakubíček and Pavel Rychlý

Natural Language Processing Centre Faculty of Informatics, Masaryk University {jak,pary}@fi.muni.cz

Lexical Computing Brno, Czech Republic {milos.jakubicek,pavel.rychly}@sketchengine.eu

**Abstract.** In this paper we present an extension of the current distributional thesaurus as available in the Sketch Engine corpus management system towards multi-word units. We explain how multi-word sketches are used to generate multi-word unit candidates, thus preserving access to the underlying corpus texts. Finally we present sample results on the British National Corpus and discuss future development as well as difficulties in evaluation.

**Keywords:** text corpus, Sketch Engine, MWE, multi-word expressions, thesaurus

## 1 Introduction

This paper elaborates on a new development implement in Sketch Engine, a leading corpus management system [3], focusing on making a distributional thesaurus for multi-word units available. Since 2006 Sketch Engine features thesaurus for single-word units, calculated using the information obtained from Sketch Engine's word sketches [6]. In 2012, a extension to the word sketch concept has been introduced towards handling multi-word sketches [4]. Since then, the single-word thesaurus basically waited to catch up the multi-word development, so here it is, finally.

In this paper we first describe Sketch Engine, introduce the concept of word sketches and the relation of word sketches to the computation of the distributional thesaurus. We continue by explaining how multi-word sketches are calculated and how they are used to derive a multi-word sketch thesaurus. We conclude by showing preliminary results in the British National Corpus.

### 2 Sketch Engine

Sketch Engine is a leading text corpus management system which as of 2019 includes several hundreds of preloaded corpora, monolingual as well as parallel

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019, pp. 143–147, 2019. © Tribun EU 2019

ones, available to its users, who can also create their own corpora, have them annotated (part-of-speech tagged, lemmatized etc.) and contrast them against the preloaded ones. The preloaded corpora typically come from the web and are very large. In 2010, Sketch Engine started the so-called TenTen series of web corpora [1], aiming at building a corpus of ten billion words (10<sup>10</sup>, thus "TenTen") for as many languages as possible.

Targeting ten billion words was not a random choice: by 2010 we had a corpus of that size for English and it clearly showed that it allows many of the Sketch Engine features that work well with a one billion word corpus and single-word units, to work well also on multi-word units. Also, given the Zipfian distribution observed in natural language, it was clear that making the corpora bigger is the only possible way that would allow us to research further on the issues of multi-word expressions.

#### 3 Word sketches

A word sketch is a short summary of a word's collocational behaviour from the perspective of individual grammatical relations (noun's modifier, verb's subject etc.), as can be seen from the example given in Figure 1.

,≓		x Ø	.≓	Ĩ	) ×	₽		x Ø			x Ø
modifiers of "account"			nouns modified by "account"			verbs with "account" as object			verbs with "account" as subject		
bank	88,271		holder	10,883		open	26,686		belong	955	
bank accou			account holder	-		create	50,014		accounts bel		
twitter Twitter acco	35,635 ount		deficit current account	7,635 t deficit		delete	5,276		balance account bala	348 nces	
email email accou	24,059		balance account balance	9,838		register	5,661		differ accounts diff	528	
	26.077		receivable	.e 3.912		access	7,391		unbanned	er 298	
user accour			accounts receiv	-/		manage	11,442		to have the a		
checking checking ad	10,970		executive Account Execut	8,498		check	5,122		open account oper	1,295	
facebook	13.512		manager	21.579		close	5,161		exist	960	
Facebook a			Account Manag			activate	2,851		into account		
<b>detailed</b> a detailed a	13,386 account of		password account passw	3,362 ord		link note that E	4,179 Education		expire account has	322 expired	
<b>paypal</b> PayPal acco	8,434 ount		surplus current account	2,371 t surplus		take	48,517		allow account allow	1,716 vs you	

Fig. 1: An example of a word sketch for the English noun *account*.

Each word sketch item is a triple consisting of the headword, the grammatical relation and the collocate. As such a word sketch is basically a dependency syntax graph, calculated using a hybrid rule-based and statistical approach. The backbone word for computing word sketches represents a hand-written word

sketch grammar, which selects collocation candidates using the corpus query language (CQL, [2]).

A sketch grammar typically makes heavy use of regular expressions over morphological annotation of the corpus to select syntactically viable collocation candidates. These candidates are subsequently subject to statistical scoring using a word association score. LogDice is used as the association metric in Sketch Engine as it was proven to be scalable across corpora of different sizes and produces scores comparable across corpora too [5].

In [4], an extension to the word sketch formalism has been presented which allowed the users to obtain a word sketch for multi-word units. That development took a very flexible approach towards multi-word expressions: any two or more words interlinked with word sketch relation formed a multiword expression for which word sketches were calculated. This method has one strong advantage, namely that it accounted well for discontinuous multiword expressions (because word sketch relations may catch rather long distance dependencies) and one obvious disadvantage that the words hat to be linked in the sketch grammar, therefore less developed sketch grammars did not provide so many multi-word expressions.

toct (noun)	Alternative PoS: verb (freq	: 941,372)
1C31 enTen1	Alternative PoS: <u>verb</u> (freq <b>`en [2012] freq = <u>1,915,482</u></b>	(147.70 per million)

Lemma	Score Fre	eq	
testing	0.520 558	3,727	requirement operation
<u>assessment</u>	0.410 640	),347	report requirement operation evaluation model application
<u>analysis</u>	0.399 1,196	6,660	evaluation model management
procedure	0.382 1,311	.,372	evaluation analysis technology
<u>study</u>	0.380 3,090	,402	
<u>method</u>	0.373 2,760	),051	
application	0.366 3,171	.,582	offocess of the
program	0.365 6,442	2,955	strategy rue product exercise Testinc
<u>datum</u>	0.362 3,165	5,540	assessmenevel record the dass and a
evaluation	0.360 468	3,130	treatment measure nronram research
<u>model</u>	0.357 2,557	',538	feature exam service program in the degree course mothod
<u>training</u>	0.354 2,486	6,409	work activity project plan datum moment
<u>research</u>	0.354 3,171	.,715	
examination	0.352 375	5,991	result check technique training
<u>requirement</u>	0.349 1,734	,482	practice performance
<u>exam</u>	0.349 373	8,769	development solution
<u>review</u>	0.348 1,803	3,362	alippicara

Fig. 2: An example of the thesaurus for the English noun *test*.

#### 4 Thesaurus

On top of the word sketches a distributional thesaurus has been part of Sketch Engine since 2006, facilitating an efficient algorithm which was tracktable

on multi-billion word corpora [6]. The thesaurus is using word sketches for computing the similarity score: it basically compares word sketch collocations for every pair of words in the corpus and the similarity relates to the fraction of shared collocates between these two words, taking the collocation weights as given by logDice into account.<sup>1</sup> A thesaurus screenshot can be found in Figure 2.

The new multi-word extension of the thesaurus uses the multi-word sketches as its backbone. The calculation starts by dumping the whole word sketch database and discovering multi-word sketches (i.e. two and more words connected with a word sketch relation) with a minimum frequency of 100 (less frequent items are unlikely to have any salient thesaurus items). These items form a new multi-word thesaurus lexicon in addition to single-word items, and are subject to the normal thesaurus calculation. In Figure 1 we show a sample multi-word entry obtained from the (by current measures, rather small) British National Corpus [7].

score	frequency	item
1.00	755	kohl chancellor helmut
0.88	1790	kohl helmut
0.75	7307	kohl chancellor
0.34	606	kohl
0.20	140	mitterrand president
0.18	536	chancellor kohl
0.17	340	bush president us
0.17	153	bush us president
0.17	20	re-unification
0.17	116	bush us
0.16	370	bush george president
0.16	283	bush president george
0.16	116	clinton president

Table 1: Thesaurus items for the phrase "kohl helmut chancellor" on the BNC.

#### 5 Conclusions and Future Work

In this paper we have presented an extension allowing the Sketch Engine's thesaurus to be applied to multi-words. The biggest advantage of the approach is that (by relying on the multi-word sketches), it makes very few assumptions

<sup>&</sup>lt;sup>1</sup> See https://www.sketchengine.eu/documentation/statistics-used-in-sketchengine/ for an exact formula.

on the form of the multi-word expressions: any two or more words are allowed as long as they are connected through a sketch grammar relation.

The results immediately indicate further space for improvement: as of now the order of the words in the multi-word expression matters, thus "kohl helmut chancellor" is different (but most similar) to "kohl chancellor helmut", which is not very user-friendly. For production use the phrases will be handled disregarding the word order. The most challenging aspect of the thesaurus development is its evaluation though, whether it concerns single-words or multi-words. Assessing word's similarity is a very difficult task for humans and therefore it is very difficult to obtain reliable evaluation datasets feature high inter-annotator agreement. Nevertheless this is a topic that we want to focus on in the future.

**Acknowledgements** This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin infrastructure LM2015071 and OP VVV project CZ.02.1.01/0.0/0.0/16\_013/0001781. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

#### References

- 1. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family. International Conference on Corpus Linguistics, Lancaster (2013)
- Jakubíček, M., Rychlý, P., Kilgarriff, A., McCarthy, D.: Fast syntactic searching in very large corpora for many languages. In: PACLIC 24 Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation. pp. 741–747. Tokyo (2010)
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. Lexicography 1 (2014). https://doi.org/http://dx.doi.org/10.1007/s40607-014-0009-, http://dx.doi.org/ 10.1007/s40607-014-0009-9
- 4. Kilgarriff, A., Rychlý, P., Kovár, V., Baisa, V.: Finding multiwords of more than two words. Proceedings of EURALEX2012 (2012)
- Rychlý, P.: A lexicographer-friendly association score. Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN pp. 6–9 (2008)
- Rychlỳ, 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)
- 7. Aston, G., Burnard, L.: The BNC handbook: Exploring the British National Corpus with SARA. Edinburgh University Press (1998)