Development of HAMOD: a High Agreement Multi-lingual Outlier Detection dataset

Miloš Jakubíček[†], Emma Romani[‡], Pavel Rychlý[†], Ondřej Herman[†]

†Natural Language Processing Centre
Faculty of Informatics, Masaryk University, Brno, Czechia
{jak,pary,xherman1}@fi.muni.cz

†Lexical Computing
Brno, Czechia
{milos.jakubicek,pavel.rychly,ondrej.herman}@sketchengine.eu

‡Università degli studi di Pavia
Pavia, Italy
emma.romani01@universitadipavia.it

Abstract.
In this paper we describe further development of a High Agreement Multi-lingual Outlier Detection dataset (HAMOD) outlier that is used for the purpose of evaluation of automatic distributional thesauri. We briefly introduce the task and methodological motivation for developing such a dataset, then we present the current status of the dataset and related tools as well as results measured on the dataset so far (both in terms of agreement rates and thesauri eveluation). Finally we discuss future developments of HAMOD.

Keywords: HAMOD · Distributional thesaurus · Outlier detection · Word embeddings · Sketch Engine

1 Introduction and motivation

This paper presents new developments of the HAMOD dataset. HAMOD stands for an acronym of *High Agreement Multi-lingual Outlier Detection*, a dataset for exercising the outlier detection task that aims at high inter-annotator agreement. Outlier detection is a task where a human or machine is presented with a set of words (in our case 9), out of which one is a so called *outlier*: a word that "doesn't fit" to the others.

In [1] it was argued that outlier detection is (unlike the intrinsic evaluation based on similarity judgements) a reliable method for evaluating automatic distributional thesauri. A distributional thesaurus is generally a mapping of pairs of words to a numeric similarity score (or conversely, a dissimilarity score, i.e. a distance) yielding in the first place a list of most similar words for a given word. There are several methods for calculating a distributional thesaurus, such as using word sketches in Sketch Engine [2] or using a vector space model

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021, pp. 177–183, 2021.

(word embeddings) (see e.g. [3]). The real difficulty for any comparison and further development of these methods is that a reliable evaluation methodology is currently missing: a directly intrinsic evaluation suffers from extremely low inter-annotator agreement. For this reason we started developing HAMOD in 2019 and continuously expand the dataset both in terms of number of languages and number of exercises.

In further text we describe the dataset itself, thesauri that we used for evaluation so far and our plans for further development.

2 Sketch Engine and the word sketch-based thesaurus

Sketch Engine [4] is a leading text corpus management system which as of 2021 includes several hundreds of preloaded corpora as well as corpus-building functionalities available for regular end users. The preloaded corpora typically come from the web and aim at targeting multi-billion size. In 2010, Sketch Engine started the so-called TenTen series of web corpora [5], aiming at building a corpus of ten billion words (10^{10}, thus "TenTen") for as many languages as possible.

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.

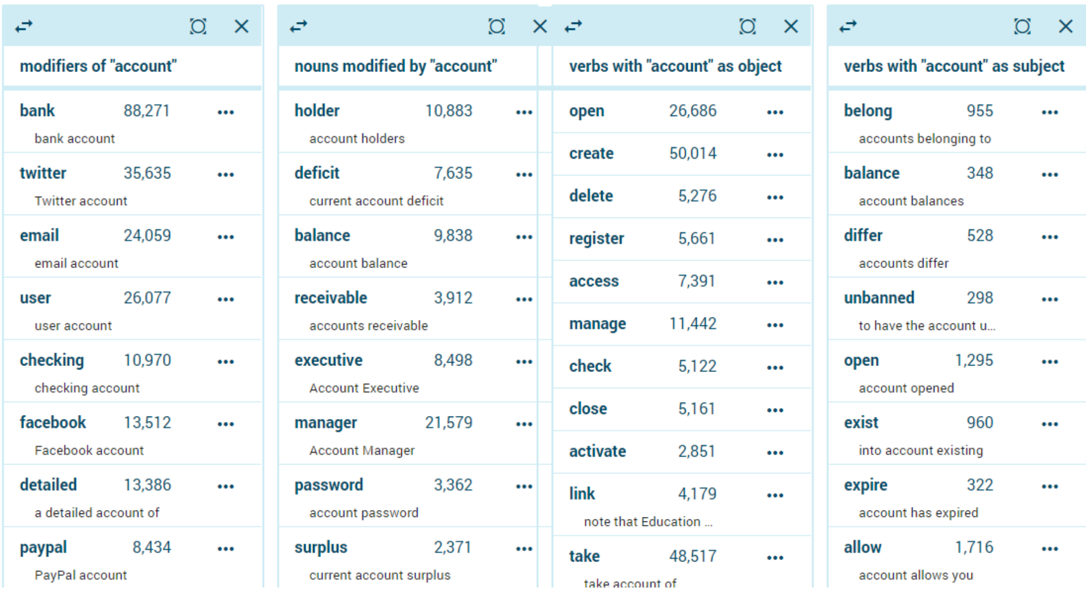

Fig. 1: An example of a word sketch for the English noun *account*.

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, [6]).

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 [7].

Word sketches make it possible to automatically derive a distributional thesaurus by calculating similarity of word sketch contexts: for each word, we look at which other words share most collocates (in the same grammatical relations).

To compute a similarity score between word w_1 and word w_2, we compare w_1 and w_2's word sketches in this way:

- find all the overlaps, i. e. where w_1 and w_2 share a collocation in the same grammatical relation, e. g.: $(beer/wine, OBJECT\_OF, drink)$, where the association score > 0,
- let ws_{w_1} and ws_{w_2} be the set of all word sketch triples $($*headword*, *relation*, *collocation*$)$ for w_1 and w_2, respectively, where the association score > 0,
- let $ctx(w_1) = \{(r,c)|(w_1,r,c) \in ws_{w_1}\}$,
- let AS_i be the association score of a word sketch triple (logDice),
- then the distance between w_1 and w_2 is computed as:

$$Dist(w_1,w_2) = \frac{\sum\limits_{(r,c)\in ctx(w_1)\cap ctx(w_2)} AS_{(w_1,r,c)} + AS_{(w_2,r,c)} - \frac{\left(AS_{(w_1,r,c)} - AS_{(w_2,r,c)}\right)^2}{50}}{\sum_{i\in ws_1} AS_i + \sum_{i \in ws_2} AS_i}$$

The term $(AS_i - AS_j)^2/50$ is subtracted in order to give less weight to shared triples, where the triple is far more salient with w_1 than w_2 or vice versa. We find that this contributes to more readily interpretable results, where words of similar frequency are more often identified as near neighbours of each other.

A thesaurus screenshot from Sketch Engine can be found in Figure 2.

3 Thesaurus built from word embeddings

Another method, or rather a whole paradigm, that can be used for deriving an distributional thesaurus, is based on calculating a vector representation for each word in a corpus (so called word embedding) and using the distances between individual word vectors as a measure of words' (dis)similarity. For our experiments we used FastText [8] and Word2vec [3] to calculate word embeddings based on corpora available in Sketch Engine [9].

test *(noun)* Alternative PoS: verb (freq: 941,372)
enTenTen [2012] freq = 1,915,482 (147.70 per million)

Lemma	Score	Freq
testing	0.520	558,727
assessment	0.410	640,347
analysis	0.399	1,196,660
procedure	0.382	1,311,372
study	0.380	3,090,402
method	0.373	2,760,051
application	0.366	3,171,582
program	0.365	6,442,955
datum	0.362	3,165,540
evaluation	0.360	468,130
model	0.357	2,557,538
training	0.354	2,486,409
research	0.354	3,171,715
examination	0.352	375,991
requirement	0.349	1,734,482
exam	0.349	373,769
review	0.348	1,803,362

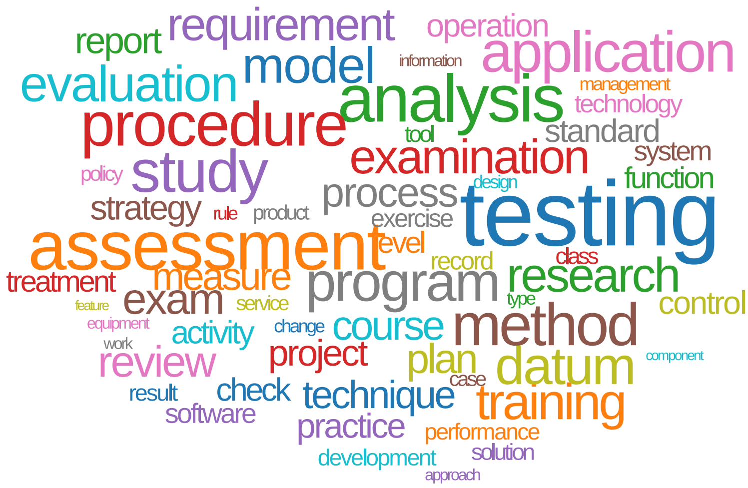

Fig. 2: An example of the thesaurus for the English noun *test*.

Unlike the corpora used for the word-sketch based thesaurus, corpora used for training word embeddings do not need to be part-of-speech tagged or lemmatized, on the other hand our preliminary observations showed that much larger datasets are required. This observation is to be expected and represents a typical data richness vs. data size trade-off.

4 Building HAMOD

In 2019 we started building HAMOD, initially on a set of three languages (English, Czech and Slovak). Currently, four other languages were added (Estonian, French, German and Italian) and we plan to expand the dataset further on. New languages are added by translating from English but where the translation results into ambiguities in the target language, we adjust the exercise set accordingly. Thus the dataset is not strictly a parallel one but a comparable one. Each exercise set of HAMOD contains 8 inliers, i.e. words that are part of a semantic category or together define a topic an, and 8 outliers. In each exercise all inliers and one outlier is presented, thus we have 8 exercises available for each such exercise set.

Since key aspect of HAMOD is the high agreement, we developed a simple web interface for exercising the outlier detection tasks by human evaluators. We aim at having at least 10 independent evaluations for each exercise and each human evaluator should be presented with an exercise set only once (i.e. never multiple times with different outliers where we could reuse the information from previous run), therefore we need 80 evaluators at minimum for each language. After completing the whole exercise, we present the evaluator with an overall success score, but do not disclose individual discrepancies.

A screenshot from the web inteface used for evaluation is provided in Figure 3. In each turn of the exercise, evaluators select the outlier, or may skip the turn if they are unsure. Currently HAMOD contains 38 complete exercise sets and the target size for all languages is 100.

5 Evaluation

Initial evaluation of the inter-annotator agreement for Czech and Estonian shows very promising results as it exceeds 90 % of absolute raw agreement (chance-correction does not play a big role: with 10 annotators and 8 options chance agreement is $\frac{1}{8}^{10} < 10^{-10}$). Detailed agreement figures for both languages are provided in Table 1.

Table 1: Inter-annotator agreement for languages included in HAMOD. A success run means an excercise where all sets where correctly fulfilled by an evaluator.

Language	**Success runs**	**All runs**	**Agreement**
Czech	2,082	2,150	0.97
Estonian	3,285	3,525	0.93

Evaluation of two distributional thesauri by means of overall accuracy (where the outlier was correctly identified) and outlier position percentage (OPP, average percentage of the right answer) is provided in Table 2. We used the czTenTen12, deTenTen13, enTenTen13, frTenTen12, itTenTen16, skTenTen11 [5] and EstonianNC 2017 [10] corpora available in Sketch Engine. For a detailed description of the evaluation, see [1].

The evaluation of the thesauri is clearly just a starting point but it already shows that none of the variants (thesaurus based on word sketches and thesaurus based on word embeddings) outperforms the other one for all languages.

6 Conclusions and future development

In this paper we have described recent developments of the HAMOD dataset. We argued why such a dataset is necessary for further development, evaluation and comparison of distributional thesauri and we have discussed the current status of the dataset. We plan to further expand the dataset to reach 100 exercises sets and cover more languages (EU languages in the first place) while continuously monitoring the inter-annotator agreement and adjusting the dataset accordingly to maintain high agreement. So far the discriminative power of the dataset (i.e. its ability to discover differences between individual thesaurus types) is maintained as well but we are aware of the fact that at

Table 2: Comparison of a Sketch Engine-based and word-embeddings-based thesaurus on the HAMOD dataset. Dataset size means number of exercises (outlier detection exercise sets) that were evaluated.

Corpus	Corpus size	Dataset size	SkE Acc	SkE OPP	Word2Vec Acc	Word2vec OPP
czTenTen12	5G	232	0.573	0.898	0.655	0.871
enTenTen13	22G	296	0.456	0.847	0.655	0.873
EstonianNC 2017	1.3G	296	0.564	0.832	0.547	0.784
deTenTen13	19G	232	0.349	0.798	0.323	0.764
frTenTen12	6.8G	232	0.276	0.744	0.427	0.768
skTenTen11	0.6G	296	0.389	0.777	0.591	0.851
itTenTen16	5.8G	296	0.453	0.856	0.581	0.869

some point of further development of the thesauri the dataset might need to be revisited if it looses its discriminative power, i.e. if it would be a task too easy for the computer. When finished the dataset will become available under a permissible Creative Commons licence in a public repository.

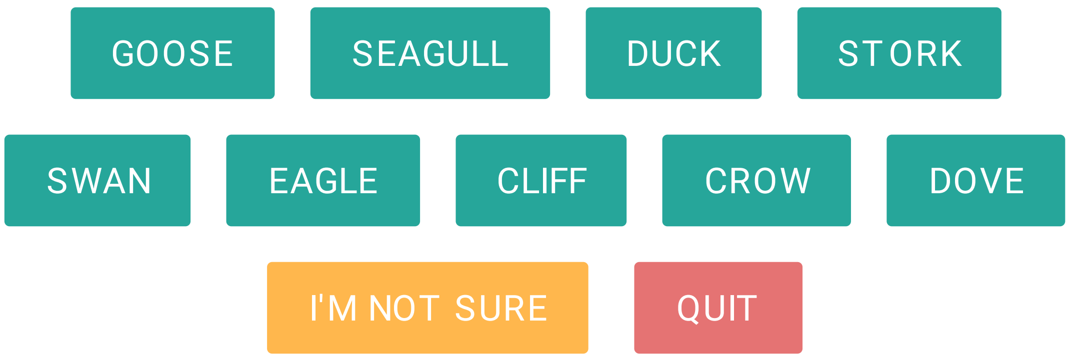

Fig. 3: A sample outlier detection exercise generated for English.

Acknowledgements. This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

References

1. Rychlý, P.: Evaluation of czech distributional thesauri. In: RASLAN 2019 Proceedings of Recent Advances in Slavonic Natural Language Processing. (2019) 137–142

2. 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
3. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. arXiv preprint arXiv:1712.09405 (2017)
4. 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)
5. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family. International Conference on Corpus Linguistics, Lancaster (2013)
6. 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, Tokyo (2010) 741–747
7. Rychlý, P.: A lexicographer-friendly association score. RASLAN 2008 Proceedings of Recent Advances in Slavonic Natural Language Processing (2008) 6–9
8. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5** (2017) 135–146
9. Herman, O.: Precomputed word embeddings for 15+ languages. RASLAN 2021 Proceedings of Recent Advances in Slavonic Natural Language Processing (2021)
10. Koppel, K.: Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele. Tartu Ülikooli Kirjastus (2020)