Czech Dataset for Semantic Similarity and Relatedness

Miloslav Konopík and **Ondřej Pražák** and **David Steinberger**
NTIS – New Technologies for the Information Society,
Department of Computer Science and Engineering,
Faculty of Applied Sciences, University of West Bohemia, Technická 8, 306 14 Plzeň
Czech Republic
konopik@kiv.zcu.cz
ondfa@ntis.zcu.cz
fenic@students.zcu.cz

Abstract

This paper introduces a Czech dataset for semantic similarity and semantic relatedness. The dataset contains word pairs with hand annotated scores that indicate the semantic similarity and semantic relatedness of the words. The dataset contains 953 word pairs compiled from 9 different sources. It contains words and their contexts taken from real text corpora including extra examples when the words are ambiguous. The dataset is annotated by 5 independent annotators. The average Spearman correlation coefficient of the annotation agreement is $r = 0.81$. We provide reference evaluation experiments with several methods for computing semantic similarity and relatedness.

1 Introduction

Computational methods for automatic assessment of semantic similarity of words significantly changed NLP in recent years.

Evaluation datasets, such as the one introduced in this work, play a crucial role in the development of methods for computing semantic similarity of words. The evaluation datasets consist of word pairs, with associated values of their semantic similarity or relatedness (for example "rooster" and 'hen'" → "high"),. The automated methods for computing semantic similarity are then evaluated according to how much the computed output of a method agrees with the human judgment present in the evaluation dataset.

The selection of the words in the dataset is very important. Therefore we introduce four different methods of selecting the word pairs in our dataset. Such diversity of sources ensures that no evaluated method can by a chance or by purpose focus on a certain way to prepare the training data to achieve better results in the evaluation.

Similar	Related
car – automobile	car – road
rooster – cock	rooster – hen
puck – biscuit	puck – hockey
water – irrigate	irrigate – field

Table 1: Examples of semantically similar and related words.

Semantic relations of words can be perceived as *semantic similarity* or *semantic relatedness*. *Semantic similarity* of words indicates how much of the meaning the words share. The higher similarity of the word, the higher probability is that the words can be replaced one with another in a sentence without changing the meaning of the sentence. On the other hand, *semantic relatedness* describes how much are the words related in meaning. The higher relatedness the higher chance that the words appear in semantically related texts. Some examples are shown in Table 1. Here we can see that similarity implies the same parts of speech for both words whereas relatedness can be high even for words with different parts of speech.

2 Related Work

Several evaluation datasets appeared in the last decade in relation to the increasing interest in the computational models for semantic similarity. The datasets exist primarily for English, however, there are datasets for other languages as well.

The *Rubenstein-Goodenough* (Rubenstein and Goodenough, 1965) dataset was introduced already in 1965. It consists of 65 pairs of regular English words. The similarity is given on scale from 0 to 4. The inter-annotator agreement for the dataset is $r = 0.85$ of Spearman correlation.

Proceedings of Recent Advances in Natural Language Processing, pages 401–406,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_053

The *Wordsim* (Finkelstein et al., 2002) dataset was designed with the purpose of enlarging the dataset existing at that time. It contains 353 pairs of nouns. It includes 30 words from the *Rubenstein-Goodenough* dataset and 82 pairs where at least one of the words is not contained in the *Wordnet*. The dataset is divided into similarity and relatedness subsets. The former contains 203 word pairs and the later 252 pairs (102 pairs are shared). 16 annotators participated in creation of the dataset and they evaluated the similarity on a 0–10 discrete scale. The inter-annotator agreement reached $r = 0.72$.

The *MTurk* (Radinsky et al., 2011) dataset consists of 280 word pairs generated from *New York Times* papers. The words must occur in the *DBPedia* database (Lehmann et al., 2014). The Amazon's Mechanical Turk service was used to annotate the semantic relatedness scores (1–5 scale).

The *Rare words* (Luong et al., 2013) dataset consists of words that occur rarely in common texts (in this case, Wikipedia articles). The words are divided in 5 groups according to their frequency in Wikipedia. To exclude foreign (non-English) words, all words are checked against the *WordNet* database (Miller, 1995). For generating the pairs, the second words are taken from *WordNet*. Some relation or relations are selected and the second word is found (e.g. the second word must be a hyponym or a hypernym of the first word). In this way, 2034 word pairs were constructed and consequently annotated via the Amazon's Mechanical Turk service (0–10 scale).

The *MEN* (Bruni et al., 2014) dataset contains words used for tagging images. The dataset is primarily designed to evaluate multi-modal computational models, however, it can be used as well for text data only. The word pairs were generated randomly. In order to avoid majority of pairs with low semantic relatedness, the pairs were scored by the *HAL* semantic model (Lund and Burgess, 1996) and some of the pairs with low scores were discarded. The Amazon's Mechanical Turk service was used to annotate the semantic relatedness scores, however, the annotators were instructed to make binary decisions which of two given word pairs are more related. The relatedness scores for 3000 word pairs were computed from these binary decisions.

There are three evaluation datasets for German. The *Gur65* dataset (Gurevych, 2005) is created by translating the 65 pairs from the *Rubenstein-Goodenough* English dataset. The scores for semantic similarity were newly annotated by 24 subjects with the inter-annotator agreement of $r = 0.81$. The *Gur350* dataset (Zesch and Gurevych, 2006) consists of 350 word pairs with relatedness scores assigned on a 0-4 scale by 8 annotators (iter-annotator agreement $r = 0.69$). The *ZG222* dataset (Zesch and Gurevych, 2006) contains 222 word pairs annotated by 21 subjects on a 0-4 scale (iter-annotator agreement $r = 0.49$).

A cross-lingual dataset for English, Spanish, Arabic, Romanian languages is described in (Hassan and Mihalcea, 2009). The dataset is created by translation from two English datasets into Spanish, Arabic and Romania. The semantic relatedness scores are taken directly from the English datasets.

The only dataset of semantic similarity scores for Czech is presented in (Krčmář et al., 2011). The dataset consists only of 55 out of 65 word pairs translated from the *Rubenstein-Goodenough* dataset. The 10 pairs were left out due to problems with translation. 55 pairs are insufficient for proper evaluation since the confidence intervals for Spearman correlation coefficient[1] are very wide at these low counts.

3 Dataset Design

3.1 Czech Language

The presented dataset is created in the Czech language. We begin by introducing the very basics of Czech. Czech belongs into the Indo-European, West Slavic language family. Czech is a synthetic language with a high ratio of morphemes per word. The morphology of the Czech language is rich and highly irregular. Czech syntax follows the subject verb object sentence structure, however, the word order is frequently altered to stress out certain words in the sentence.

3.2 Word Pairs Selection

In order to obtain high quality dataset, we use four methods for selecting the words for the word pairs. The first method extracts word pairs used in existing English datasets. The English pairs are translated into Czech and included in the Czech dataset. In the second method, the pairs are extracted from the translation tables for machine translation. The third method of pair generation is based upon the

[1] Spearman correlation coefficient is explained in Section 4.1

Method	Source	Count
Translation	RG	46
	Wordsim	205
	MTurk	97
	MEN	121
	Rare Words	85
Translation tables		108
SCIO		118
Own		173
Total		**953**

Table 2: The composition of the new Czech semantic dataset.

bag	brašna	0.039003
bag	batoh	0.013740
bag	balíček	0.005873
bag	balík-1_^(předmět)	0.003546
bag	balení_^(*3it)	0.001995
bag	bago	0.001884
bag	aktovka	0.001662
bag	airbag	0.001662
bag	bags	0.001551
bag	balit_:T	0.000997

Table 3: A snapshot of a translation table. The words are lemmatized.

SCIO language quiz [2]. The rest of the pairs were invented by the annotators. The counts for each method are given in Table 2. We explain the methods in more detail in the following text.

Translating the existing datasets. The annotators translated to Czech randomly selected pairs from the following English datasets: *RG*, *MTurk*, *MEN* and *Rare words*. The annotators were instructed to refrain from using any translation service or translation dictionary. Instead, they used their knowledge or an explanatory dictionary. In this way, they were forced to think about the translation in terms of the original pair, not in terms of individual words. The translations were prepared by two annotators for each pair and the different translations were discarded. We also discarded the translations where the original English word can be translated only as a phrase (e.g. "seafood" → "plody moře"). The resulting counts of pairs for each dataset are shown in Table 2. The word pairs in the used English datasets employ different methods of pair selection. Thus, we obtain a multi-source list of pairs just by taking some pairs from each of them.

Extraction from Translation Tables. This method is based upon the bilingual pivoting technique (Bannard and Callison-Burch, 2005). In this technique, bilingual parallel corpora are first aligned on the word level. Next, the pivots are found by looking for foreign words that have different translations. Finally, the different translations are scored according to the alignment probability and frequency in the corpus. The most probable translations for a given pivot are considered as equal in meaning. We simplify the procedure by using the translation tables from a machine translation system. We take a foreign word and look for different translations of the word in the translation tables – see example in Table 3: "brašna" means bag and "bathoh" means backpack whereas "balit_:T" means to pack. To select semantically similar and dissimilar word pairs we always take the first record in the translation table but we randomly select the second record. The similar words tend to be at the top of the translation table, however, at the end the words tend to be somehow related but fairly dissimilar. To generate the translation tables, we use the Moses system (Koehn et al., 2007) and CZENG corpus (Bojar et al., 2011).

SCIO. SCIO tests are used in the Czech Republic for testing the students' general knowledge for university administration exams. The tests include the task to select a most similar, related or antonymous word for a given group of words. We randomly sampled from the groups of words to generate the pairs.

Own Inventions. As the last method, the annotators were asked to invent their own pairs. The annotator were instructed to invent similar, related, antonymous and unrelated pairs.

3.3 Structure of the Dataset

The dataset is structured in records of 8 values:

1. Word 1 – the first word of the pair – e.g. "kohout" (rooster).

2. Word 2 – the second word of the pair – e.g. "slepice" (hen).

[2] SCIO tests are used in the Czech Republic for testing the students' general knowledge for university administration exams.

3. Similarity – discrete scale from 0 to 5 – e.g. 3 for "kohout" (rooster) and "slepice" (hen).

4. Relatedness – discrete scale from 0 to 5 – e.g. 5.

5. Context 1 – Context for the word 1 – e.g. "**Kohout** běhal po dvoře" (The **rooster** ran in the yard).

6. Context 2 – Context for the word 2 – e.g. "**Slepice** sedí na vejcích." (A **hen** is sitting on eggs).

7. Ambiguity – An example of ambiguity in case one of the words is ambiguous: "Je otevřen odběrový **kohout**." (The **tap** is open).

8. Common context – A block of text where both words appear together – e.g. "**Slepice** a **kohout** běhali po dvoře" (The **hens** and the **rooster** ran in the yard).

All examples used in the dataset are taken from the SYN corpus (Hnátková et al., 2014). The examples were found via the `Korpus.cz` page.

4 Dataset Annotation

The word pairs in the dataset were annotated by 5 annotators. Two of them were high school teachers of Czech, the others were students. All received oral instructions and a simple annotation manual with examples.

4.1 Inter-annotator Agreement

We use the Spearman correlation coefficient – see Equation 4.1 to compute the inter-annotator agreement of all 5 annotators:

$$r(\mathbf{x}, \mathbf{y}) = 1 - \frac{6 \sum_{i=0}^{n} (r_{x_i} - r_{y_i})^2}{n \times (n^2 - 1)}, \qquad (1)$$

where r_{x_i} a r_{y_i} are two ranks of scores x_i and y_i for the i-th pair and n is the number of word pairs (in our case $n = 953$).

The resulting average Spearman correlation coefficient for all 5 annotators is $r = 0.81$. All of the 5 annotators annotated all the words in the dataset. The correlation is computed for all annotators and it is averaged.

Dataset	Similarity	Relatedness
RG65	90.16	88.56
WS353*	88.70	71.12
MTurk	66.90	70.94
MEN	82.73	87.58
Rare Words	66.25	56.05

Table 4: Spearman correlation coefients for the similarity and relatedness scores between the new Czech corpus and the English corpora. The values are multiplied by 100 for better oriantation. RG65 is the Rubenstein-Goodenough dataset, *WS353* is the Wordsim dataset. * For the WordSim dataset, the similarity corelation is computed on the similarity part of the corpus and the relatedness correlation of the relatedness part.

4.2 Inter-dataset Agreement

Table 4 shows the correlation of similarity and relatedness scores between the new Czech corpus and the English corpora. The Spearman correlation coefficient is computed for the scores of all the translated words and the original scores in the corresponding English corpora. For the WordSim dataset, the similarity correlation is computed on the similarity part of the corpus and the relatedness correlation on the correlation part.

The resulting correlation is high except for the MTurk and the Rare Words datasets. The low correlation for the Rare Words dataset can be explained by the nature of the words in the dataset. The words are rare and difficult to translate. Even when translated using the explanatory dictionaries the translation could not be perfected. The meaning is slighlty different for most of the translated words. It is hard to explain the correlation for the MTurk dataset since no inter-annotator agreement is published for this dataset. For the Rubenstein-Goodenough and the WordSim datasets the obtained correlation coefficients are very close to the published inter-annotator agreements.

4.3 Scores Distribution

Table 5 shows the distribution of similarity and relatedness scores across the dataset. We can observe that trends for similarity and relatedness are reversed. There is a little number of very similar word pairs with score 5 and the counts go up with decreasing similarity (in average). The trend is reversed for semantic relatedness. It is quite expected since when two words are similar then they

Score	Similarity	Relatedness
0	27,60% (263)	1,36% (13)
1	21,20% (202)	8,08% (77)
2	13,33% (127)	11,75% (112)
3	18,36% (175)	14,06% (134)
4	12,91% (123)	35,89% (342)
5	6,61% (63)	28,86% (275)

Table 5: Distribution of scores for similarity and relatedness.

	GloVe	CBOW	S-G	LDA
CZ–sim	50.52	54.69	58.75	40.75
CZ–rel	49.77	50.65	55.55	40.02
RG65	66.20	68.74	71.72	57.63
WS353–sim	57.21	70.85	71.58	56.82
WS353–rel	43.29	52.05	52.50	45.61
MTurk	58.35	66.14	64.83	49.92
RW	24.93	26.02	21.00	16.21
MEN	64.60	71.09	72.06	55.76

Table 6: Reference Evaluation Experiments. *CZ–sim* and *CZ–rel* are results for the new Czech dataset computed for the similarity and relatedness scores. RG65 is the Rubenstein-Goodenough dataset, *WS353–sim* and *WS353–rel* are the similarity and relatedness parts of the Wordsim dataset and *RW* is the Rare words dataset. S-G stands for Skip-gram.

are also related.

5 Reference Evaluation Experiments

Table 6 shows Spearman correlation scores for several popular methods for computing semantic similarity and relatedness. The results are shown for GloVe (Pennington et al., 2014), CBOW and Skip-gram from the Word2Vec toolkit (Mikolov et al., 2013) and for LDA (Blei et al., 2003). Results for several selected English datasets are shown for comparison.

The Czech models were trained on the Wikipedia dump. For English, we have used a subset of Wikipedia articles with comparable size to the Czech corpus.

The results show that the dataset is fairly difficult comparing to its English counterparts. Only the Rare words dataset and the relatedness part of the WordSim dataset provided lower correlations. We believe that the difficulty of the dataset is caused mainly be the properties of the Czech language – see section 4.1. Given the relatively high inter-annotator agreement ($r = 0.81$), there seems to be a sufficient room for further improvements of the methods for computing semantic similarity and relatedness.

6 Conclusion

We introduce not yet another dataset but a dataset different in several areas. It is available for download at: `https://goo.gl/KctX2X`. We introduced a new technique to obtain word pairs for the corpus based upon bilingual pivoting. The obtained inter-annotator agreement of $r = 0.81$ is sufficiently high.

6.1 Distinguishing Properties of the Introduced Dataset

- Semantic similarity and relatedness is provided separately for all words.
- The word pairs are created by four different methods.
- The dataset works with word senses. Each word is considered in its sense that is given by an example.

Acknowledgments

This publication was supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports and by Grant No. SGS-2016-018 Data and Software Engineering for Advanced Applications. Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures".

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '05, pages 597–604. https://doi.org/10.3115/1219840.1219914.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022. http://dl.acm.org/citation.cfm?id=944919.944937.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček,

Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2011. Czech-english parallel corpus 1.0 (CzEng 1.0). LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University. http://hdl.handle.net/11234/1-1458.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)* 49(1-47).

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Trans. Inf. Syst.* 20(1):116–131. https://doi.org/10.1145/503104.503110.

Iryna Gurevych. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the Second International Joint Conference on Natural Language Processing*. Springer-Verlag, Berlin, Heidelberg, IJCNLP'05, pages 767–778. https://doi.org/10.1007/11562214_67.

Samer Hassan and Rada Mihalcea. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '09, pages 1192–1201. http://dl.acm.org/citation.cfm?id=1699648.1699665.

Milena Hnátková, Michal Křen, Pavel Procházka, and Hana Skoumalová. 2014. The syn-series corpora of written czech. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '07, pages 177–180. http://dl.acm.org/citation.cfm?id=1557769.1557821.

Lubomír Krčmář, Miloslav Konopík, and Karel Ježek. 2011. Exploration of semantic spaces obtained from czech corpora. In *Proceedings of the Dateso 2011: Annual International Workshop on DAtabases, TExts, Specifications and Objects, Pisek, Czech Republic, April 20, 2011*. pages 97–107. http://ceur-ws.org/Vol-706/paper24.pdf.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Chris Bizer. 2014. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal* .

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers* 28(2):203–208. https://doi.org/10.3758/BF03204766.

Minh-Thang Luong, Richard Socher, and Christopher D. Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNLL*. Sofia, Bulgaria.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM* 38(11):39–41. https://doi.org/10.1145/219717.219748.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 1532–1543. http://aclweb.org/anthology/D14-1162.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, New York, NY, USA, WWW '11, pages 337–346. https://doi.org/10.1145/1963405.1963455.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM* 8(10):627–633. https://doi.org/10.1145/365628.365657.

Torsten Zesch and Iryna Gurevych. 2006. Automatically creating datasets for measures of semantic relatedness. In *Proceedings of the Workshop on Linguistic Distances*. Association for Computational Linguistics, Stroudsburg, PA, USA, LD '06, pages 16–24. http://dl.acm.org/citation.cfm?id=1641976.1641980.