Lecture Notes in Artificial Intelligence
Subseries of Lecture Notes in Computer Science

The LNAI series reports state-of-the-art results in artificial intelligence research, development, and education, at a high level and in both printed and electronic form. Enjoying tight cooperation with the R&D community, with numerous individuals, as well as with prestigious organizations and societies, LNAI has grown into the most comprehensive artificial intelligence research forum available.

The scope of LNAI spans the whole range of artificial intelligence and intelligent information processing including interdisciplinary topics in a variety of application fields. The type of material published traditionally includes

- proceedings (published in time for the respective conference)
- post-proceedings (consisting of thoroughly revised final full papers)
- research monographs (which may be based on PhD work)

More recently, several color-cover sublines have been added featuring, beyond a collection of papers, various added-value components; these sublines include

- tutorials (textbook-like monographs or collections of lectures given at advanced courses)
- state-of-the-art surveys (offering complete and mediated coverage of a topic)
- hot topics (introducing emergent topics to the broader community)

In parallel to the printed book, each new volume is published electronically on SpringerLink.

Detailed information on LNCS can be found at
www.springer.com/lncs

Proposals for publication should be sent to
LNCS Editorial, Tiergartenstr. 17, 69121 Heidelberg, Germany
E-mail: lncs@springer.com

ISSN 0302-9743
ISBN 978-3-319-45509-9

Lecture Notes in Artificial Intelligence
Lecture Notes in Computer Science

springer.com

Sojka et al. (Eds.)

LNAI 9924

Text, Speech, and Dialogue

TSD

Petr Sojka · Aleš Horák
Ivan Kopeček · Karel Pala (Eds.)

Text, Speech, and Dialogue

19th International Conference, TSD 2016
Brno, Czech Republic, September 12–16, 2016
Proceedings

Springer

WordSim353 for Czech

Silvie Cinková(✉)

Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Charles University, Malostranské náměstí 25, Praha 1, Prague, Czech Republic
cinkova@ufal.mff.cuni.cz
http://ufal.mff.cuni.cz

Abstract. Human judgments of lexical similarity/relatedness are used as evaluation data for Vector Space Models, helping to judge how the distributional similarity captured by a given Vector Space Model correlates with human intuitions. A well established data set for the evaluation of lexical similarity/relatedness is WordSim353, along with its translations into several other languages. This paper presents its Czech translation and annotation, which is publicly available via the LINDAT-CLARIN repository at hdl.handle.net/11234/1-1713.

Keywords: Word similarity · Word relatedness · Czech · WordSim353 · Language resource

1 Introduction

1.1 Human Judgments of Lexical Similarity/Relatedness in NLP

Recent years have seen a substantial interest in vector space modeling applied to lexical similarity or lexical relatedness[1], also in multilingual terms. A number of human judgment datasets have been created to evaluate the available semantic metrics to make sure that the metrics would simulate the human lexical similarity/relatedness reasoning. One of the first and best-known datasets of this kind is WordSim353 [3]. It has also been translated into several other languages: Arabic, Spanish, Romanian [2], and most recently also to Russian, Italian, and German [1].

To the best of our knowledge, there has not been any WordSim353 translation to Czech so far – therefore we have created one. In the following sections, we will describe the resource as well as the methodology we used, along with the results.

I thank Jan Hajič and Jana Straková for allowing me to use the Czech translations of WordSim353 they had gathered earlier.

[1] See Sect. 1.3 for a terminological clarification.

© Springer International Publishing Switzerland 2016
P. Sojka et al. (Eds.): TSD 2016, LNAI 9924, pp. 190–197, 2016.

1.2 The Original WordSim353

The original WordSim533 was created as an evaluation dataset for context-based keyword search [3]. It consists of 353 noun pairs, some of which are multiword expressions. It includes an older lexical resource – Miller and Charles' word list [4] of 30 noun pairs. The noun pairs represent diverse degrees of lexical similarity - from totally unrelated concepts to frequently co-occurring words and synonyms or antonyms in both a narrow and a broad sense. The dataset is partially related to the English WordNet. (The authors note that 82 word pairs contain at least one word not captured by WordNet.)

WordSim353 contains human similarity judgments. Human annotators were to rank the "similarity"[2] between the words in a pair, using a continuous scale from 0 ("completely unrelated") to 10 ("identical or extremely similar"). 153 word pairs were annotated by 13 annotators, the rest by 16 annotators[3].

1.3 Similarity vs. Relatedness

The semantic relations between the paired words in the WordSim353 data set are of different kinds:

1. synonyms and broader synonyms *(midday-noon, asylum-madhouse, money-cash, football-soccer)*;
2. antonyms *(smart-stupid)*;
3. hyponym + hyperonym *(bird-cock)*;
4. words frequently co-occurring in the same domain *(doctor-nurse, arrival-hotel)*;
5. parts of a multiword expression *(soap-opera)*;
6. unrelated words *(professor-cucumber)*.

Agirre et al. [6] appointed human annotators to classify the word pairs according to the semantic relation between their members, achieving a high interannotator agreement. As a result of this manual classification, WordSim353 was divided into two subsets with pairs containing mutually **similar** words and pairs containing mutually **related** words, respectively. The similarity group contained words paired by synonymy, antonymy, and hypo/hyperonymy, whereas the relatedness group contained words with relations numbered 4–6 in our list above; i.e. including words with low scores given by the original WordSim353 annotators. Agirre et al. demonstrated on this data that the similarity-relatedness distinction plays an important role in distributional semantic modeling, as the modeling of each requires different methods.

[2] Although the relation between the pair members is often referred to as "relatedness", which we also find more appropriate, the annotator instructions consistently use the word "similarity". See also Sect. 1.3.

[3] See http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/.

1.4 Multilingual WordSim353

A multilingual release of WordSim353 [2] comprises Arabic, Spanish, and Romanian. This dataset was created to evaluate vector-space representations of these languages obtained by extracting information from the interlanguage links between concept definitions in Wikipedia by means of Explicit Semantic Analysis.

The translations were obtained from native speakers of the respective languages with high proficiency in English. The translators could see the English similarity scores for each word pair and were familiar with the similarity rating task. They were instructed to provide equivalent pairs that would possibly obtain the same similarity scores as their English originals in the original task.

For Spanish, the pilot language, five independent translations were gathered, which were merged into a single selection by an adjudicator. This Spanish list of word pairs was then rescored by five independent human annotators using the same scale as the English WordSim353. The average scores for English and Spanish reached a high degree of correlation (0.86), which indicates that the translations have closely followed the original relatedness degree. Also the agreement among the translators was high (at least three annotators agreed on the same translation for a word pair). Since the pilot English-to-Spanish translation was successful, the Arabic and Romanian translations were only provided by a single translator.

1.5 Judgment Language Effect

Most recently, Leviant and Reichart [1] published a study on the effect of the judgment language on the human scoring decisions as well as on the performance of Vector Space Models, for which they built a new lexical resource – a multiligual translation of WordSim353 and SimLex999 [5]. To ensure that all their translations as well as scoring decisions followed the same policy, they rescored the English data and retranslated language versions that had already been translated for earlier lexical resources of this type. To investigate the effect of the judgment language on human decisions on lexical relatedness/similarity, Leviant and Reichart compared the interannotator agreement within each judgment language and across the judgment languages. The observed interannotator agreement was higher between than across the judgment languages, which suggests an effect of the judgment language on the human relatedness/similarity scoring. To assess the judgment language effect on the Vector Space Models, Leviant and Reichart trained two models on identical parts of the multilingual data for each language separately. Also the test data comprised identical terms across the language versions. The performance of the Vector Space Models also turned out to be language-specific. Besides, the models were also trained on the entire multilingual data. The performance of the multilingual Vector Space Models was higher than that of the monolingual ones. Leviant and Reichart draw the conclusion that both humans and Vector Space Models are affected by the judgment language, although each in a different way. The judgment-language effect appears to decrease the performance of the Vector Space Models and can be partially alleviated by multilingual setup.

Their multilingual data set comprises Italian, German and Russian as representatives of three different Indo-European language families. Two native speakers were in charge of the translation into each language. The translators were given several additional instructions beforehand to keep the translation strategy as consistent as possible. The inter-translator agreement for WordSim was 87.9 % for Russian, 91.9 % in Italian and 83.9 % in German (calculated on single words, not the entire pairs).

The translation instructions concerned the following aspects:

1. gender disambiguation (e.g. all the languages have gender-specific expressions for *cat*),
2. sense proximity (in homonymous and polysemous words, use the other word pair member to select the sense that is semantically closest to it. When the other pair member provided no hint, the translation manager selected one meaning randomly, and this interpretation was used across all three languages.)
3. POS proximity (when the part of speech of the English original was ambiguous, the translators were supposed to use an equivalent whose part of speech conforms to the other pair member).

2 The Czech WordSim353

2.1 Translation Instructions

The Czech translations were collected from four independent native speakers with high proficiency in English, largely following the translation instructions in [2]. This mainly means that i.e. multi-word expressions were not allowed and ambiguity was to be preserved whenever possible. The translation was performed long before [1] was published, hence any similarities in their translation instructions are coincidental.

The Czech translators were shown the English mean scores of each original word pair and were asked to provide a translation as close as possible in terms of the lexical similarity score: for scores between 5 and 10 the preferred equivalents should be synonyms[4] or be as closely semantically associated as possible, whereas the equivalents with low scores should make the semantic distance of the pair members as obvious as possible. On the other hand, disambiguation was not required. On the contrary, the ideal equivalent was supposed to preserve the ambiguity (knowing this was rarely possible). As Czech is rich in (almost synonymous) derivatives, the translators were also asked to use the same translation equivalent of a given word throughout the dataset. In some cases where only one of the derivatives was possible (preserving the semantic distance/proximity) in

[4] Cf. [6] and Sect. 1.3.

one pair, while making no difference in another pair, the equivalent from the more restricted pair was preferred.

An external adjudicator preliminarily merged the translations into one dataset. However, we decided to let the annotators score all sensible translation equivalents, eliminating only typos and clear mistranslations. We indexed each Czech translation equivalent with the corresponding English original.

3 Scoring Instructions

The Czech scoring instructions were obtained by translation from the English original WSim353 instructions. The document differed only by offering the respondents three different file formats for their convenience (.xls, .xlsx, and .csv). All respondents were individually contacted and their responses collected by e-mail. They were randomly chosen with respect to gender and age. The group is likely to be biased towards people with higher education and interest in language(s), but we prohibited participation to linguists (both scholars and computer scientists). The scorers did not know that the word-pair collection had an English counterpart, and naturally they were neither shown the original English word pairs nor the original English scores.

The online description of WordSim353[5] uses the word "relatedness" to describe what the annotators were rating. Nevertheless, the `instruction.txt` file containing the original annotator instructions consistently refers to "similarity", and therefore the Czech instructions also used "similarity", although "relatedness" seemed more appropriate with respect to the fact that "similarity" evokes "synonymy".

Five of the 25 Czech annotators even reacted to the word "similarity" in the instructions as soon as they had started their annotation. They suggested "relatedness" instead and asked for confirmation that "related" words should also get high ratings. When directly asked, we told them they were right but had to decide the scores according to their own intuition. We were always giving only individual answers; that is, we never intervened in the rest of the group. A few annotators complained after they had submitted their work that "there were many more similar words that were not synonyms" and hence the instructions were misleading. However, all annotators were eventually rating what they perceived as "relatedness" beyond "similarity" in the strict sense of synonymy and achieved a high pairwise correlation.

As we wanted to prevent the annotators from creating anchors based on too many close pairs that have arisen due to the translation variants, we displayed the word pairs in random order. According to the feedback delivered after submitting their responses, the annotators had not recognized that some selected words were different translations of the same English original. Pairs identified as similar to some already seen were believed to be "bogus items" to detect careless responding.

[5] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/.

4 Format

Unlike [2], we had our annotators score all translation variants of each original English word pair. This complete annotation comes as WordSim-cs-Multi, corresponding to a file named `WordSim-cs-Multi.csv`. It comprises 633 Czech word pairs (mapped on the original 353 English word pairs). From this dataset we have made a selection of Czech equivalents (WordSim353-cs, file `WordSim353-cs.csv`), whose judgments were most similar to the judgments of the English originals, compared by the absolute value of the difference between the means over all annotators in each language counterpart. In one case (*psychology-cognition*), two Czech equivalent pairs had identical means as well as confidence intervals, so we randomly selected one. In both data sets, we preserved all 25 individual scores. In the WordSim353-cs data set, we added a column with their Czech means as well as a column containing the original English means and 95 % confidence for each mean (computed by the CI function in the Rmisc R package)[6]. The WordSim-cs-Multi data set contains only the Czech means and confidence intervals. For the most convenient lexical search, we provided separate columns with the respective Czech and English single words, entire word pairs, and eventually an English-Czech quadruple in both data sets. The lexical resource also contains an .xls file with the four translation variants and adjudication before the annotation. The entire resource including the annotation and translation instructions (in Czech) is publicly available via the LINDAT-CLARIN repository at hdl.handle.net/11234/1-1713.

5 Evaluation

5.1 Correlation Between WordSim353 and WordSim353-Cs

As we were able to optimize the selection of the Czech equivalent pairs by minimizing the difference between the Czech and English mean values for each pair, the correlation between these samples does not say much about actual differences between judgment languages with respect to lexical similarity/relatedness. Nevertheless – since correlation was reported for the earlier translated WordSim-based sets, we will use the same evaluation measure: the correlation between WordSim353 and WordSim353-cs lies at 0.9 (almost identical for both Spearman's ρ and Pearson's product-moment correlation, with the confidence interval of Pearson's product-moment correlation at 0.895–0.930 and with p-value $< 2.2e-16$).

We have observed a rather symmetric distribution of the differences between the corresponding means for each language with the mean at approx. −0.1 (Czech minus English) and the standard deviation at approx. 0.9.

[6] [7] and [8].

5.2 Intertranslator and Adjudicator-Translator Agreement

We have analyzed the translations as well as the decisions of the adjudicator with Fleiss kappa. The intertranslator agreement on single words was 0.785. In 37.8% of the words (267 of 706), at least 3 annotators agreed, of which 244 reached full agreement. Pairwise, the annotators had similar agreement (0.622–0.676); there was no outlier. According to the pairwise agreement observations of the adjudicator with each translator, the adjudicator did neither seem to strongly prefer nor disprefer any translator (0.719–0.787). A manual analysis of intertranslator disagreements revealed several distinct types:

- orthographical variants, e.g., *chléb-chleba (bread), maraton-maratón (marathon), brambora-brambor (potato)*
- typing errors or omited translations (rare), e.g. *arichtektura ("arichtecture"), projekt-project, hudba-music*
- synonyms in the narrow sense (rare), e.g. *šukat-šoustat (fuck), šálek-hrnek (cup-mug)*
- synonyms - stylistic variation (formal/less formal), e.g. *chlapec-hoch (boy), auto-automobil (car), doktor-lékař (physician)*
- original Czech words vs. loan words not necessarily belonging to different registers, e.g. *teritorium-území(territory), menšina-minorita (minority), třída-avenue (avenue), katastrofa-pohroma (disaster)*
- POS - disagreement (rare), e.g. *mýdlo-mýdlová* (*soap* noun vs. derived adjective), *pít-pití* (*drink* verb vs. noun, meaning either the event of drinking or a drink), *akcie-akciový* (*share/stock* noun vs. derived adjective)
- synonyms in a broader sense, derivation difference, e.g. *vývoj-rozvoj (evolution-development), vejce-vajíčko*(*egg* vs. the diminutive form of *egg*, with partially different connotations)
- disambiguation guesses (*closet* in the pair *closet-clothes*: *wardrobe* vs. *lavatory*).

Broader synonyms sometimes differ in concept granularity and in register at the same time (e.g. *dítě-batole-miminko* for *baby*. More precisely: *child-toddler-baby*). The most interesting part of the disagreements are certainly the disambiguation guesses as well as failed or successful attempts to preserve the original ambiguity. The instructions regarding ambiguity preservation, potential similarity score matching, and consistency across the entire data set were sometimes conflicting, e.g. the first two in *alkohol-líh (alcohol-spirit)*, which occurred both in the chemical domain as well as together with alcoholic beverages, and, presumably, the association of *alkohol (alcohol)* with different types of alcoholic beverages is stronger than that of *líh* (*spirit* but not *spirits*)[7], although, technically speaking, it is the same substance, since *alkohol*, unlike *líh*, is also used as a shortcut for *alcoholic beverages*.

In a few cases, some translators did not notice (or at least not process) the fact that the pair members were parts of a multiword expression, such as

[7] A margin note: there is, though, a Czech derivative of *líh* equivalent to *spirits*, also typically used in plural: *lihoviny*, which the translators did not consider.

cell in *cell phone*. The context-independent translation *buňka (cell)* evokes no association with *telefon (phone)*, while the American English association between *cell* and *phone* is naturally tight.

In some cases, however, the English concept was so blurred, that the translation decision was necessarily random and in 8 cases led to total disagreement; e.g. *pojištění (insurance) - riziko (risk) - náchylnost (predisposition, proneness) - ručení (guarantee)* (as translations of *liability*), *zaměření (focus, specialization) - pozornost (attention) - zaostření (focus, visual accommodation, zoom) - smysl (sense, meaning)* (as translations of *focus*), *sklad (storage) - zásobit (provide, supply) - zásoba (supplies) - dobytek (livestock, cattle)* as translations of *stock*).

6 Conclusion

We have created a multiple-equivalent Czech translation and annotation of WordSim353, one of the standard datasets for evauations of semantic vector space representations. The resource is publicly available at hdl.handle.net/11234/1-1713 via the LINDAT-CLARIN repository.

Acknowledgments. This project was supported by the Czech Science Foundation grant GA-15-20031S and has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth, and Sports of the Czech Republic (project LM2015071).

References

1. Leviant, I., Reichart, R.: Separated by an Un-common Language: Towards Judgment Language Informed Vector Space Modeling. arXiv:508.00106v5 [cs.CL], 6 December 2015
2. Hassan, S., Mihalcea, R.: Cross-lingual semantic relatedness using encyclopedic knowledge. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Singapore (2009)
3. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: the concept revisited. ACM Trans. Inf. Syst. **20**(1), 116–131 (2002)
4. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Lang. Cogn. Process. **6**(1), 1–28 (1991)
5. Hill, F., Reichart, R., Korhonen, A.: Simlex-999: evaluating semantic models with (genuine) similarity estimation. [cs.CL]. arXiv:1408.3456
6. Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., Soroa, A.: A study on similarity and relatedness using distributional and wordnet-based approaches. In: Proceedings of NAACL-HLT 2009 (2011)
7. Hope, R.M.: Rmisc: Ryan Miscellaneous. R package version 1.5. https://CRAN.R-project.org/package=Rmisc
8. R Core Team: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/