SimLex-999 for Polish

Małgorzata Marciniak[1], **Agnieszka Mykowiecka**[1,2], **Piotr Rychlik**[1]
Institute of Computer Science, Polish Academy of Sciences[1]; Polish-Japanese Academy of Information Technology[2]
Jana Kazimierza 5, Warsaw, Poland; Koszykowa 86, Warsaw, Poland
{mm, agn, rychlik}@ipipan.waw.pl

Abstract

The paper addresses the Polish version of SimLex-999 which we extended to contain not only measurement of similarity but also relatedness. The data was translated by three independent linguists; discrepancies in translation were resolved by a fourth person. The agreement rates between the translators were counted and an analysis of problems was performed. Then, pairs of words were rated by other annotators on a scale of 0–10 for similarity and relatedness of words. Finally, we compared the human annotations with the distributional semantics models of Polish based on lemmas and forms. We compared our work with the results reported for other languages.

Keywords: similarity, synonymy, Polish, distributional models evaluation

1. Introduction

Distributional semantics methods are commonly used in various linguistic tasks. So creating resources to evaluate them is extremely important. The best known of these, (Gabrilovich, 2017) (Finkelstein et al., 2002), consists of 353 word pairs together with their similarity scores. Human annotators evaluated the similarity of word pairs understood as synonymy, relatedness and association of words. Another well-known resource developed for the same purpose is SimLex-999 (Hill, 2017) (Hill et al., 2015). It was created to test similarity, but here it was understood as synonymy and quasi-synonymy, which seems to be a hypernym/hyponym relation or co-hyponymy. It consists of 999 word pairs rated for this interpretation of similarity. The authors clearly excluded relatedness and association from the similarity relation. Moreover, the resource provides information on part-of-speech, abstract vs. concrete concepts, and independent measures of relatedness of pairs of words for English. Both the above resources were translated into German, Italian and Russian (Leviant and Reichart, 2015) and are available as Multilingual WS353 and Multilingual SimLex999 (Leviant and Reichart, 2017), respectively. [1]
Many authors (e.g. (Faruqui et al., 2016), (Chiu et al., 2016)) point out the disadvantages of evaluation for distributional semantics methods based on isolated tests which are not connected to a whole processing system. For languages with less developed linguistic infrastructure, such as Polish, it is important to provide resources for intrinsic evaluation too. Therefore, we decided to translate SimLex-999. As it would be interesting to compare results for synonymy and relatedness on the same resource, we rated both types of similarity relation.

2. Translation

The translation of SimLex-999 into Polish was done by three linguists, native speakers of Polish with good knowledge of English, according to the instruction published together with Multilingual SimLex999. The final translation was agreed by a fourth person with a computational linguistics background. She only resolved cases where differences in the three translations occurred, for 585 pairs in fact. As the Multilingual SimLex999 resource includes translation into Russian, a Slavic language with similar linguistic phenomena and grammar as Polish, we assumed that the most important issues were taken into account. Additionally, we wanted the translators to pay special attention to the following issues, mentioned by Leviant and Reichart in the guidelines.

- As the dataset was intended to test language models, each word should be translated into one word. This caused several problems as many English one-word terms are translated into two words in Polish. 'Sunrise' and 'sunset' in Polish are 'wschód słońca' 'zachód słońca' (shortened by all three annotators to *wschód* and *zachód*, which are polysemic and can also mean e.g. directions of the world). Moreover, many Polish verbs consist of a verb form and the reflexive pronoun *się* 'self' (in Russian a reflexive pronoun is attached to a verb, if it exists). For some verbs, this pronoun is obligatory, e.g. *śmiać się* 'laugh', while for others the existence of the reflexive pronoun changes the meaning of the verb, e.g. *słuchać* 'listen' and *słuchać się* 'obey'. The translators were ask to avoid verbs composed with the reflexive pronoun if possible.

- In Polish, as in Russian, adjectives have different forms for masculine (*stary* 'old'), feminine (*stara* 'old') and neutral (*stare* 'old') genders. In translation, we use masculine forms as dictionaries use them as lemmas.

- The translators should prefer interpretations which make the words in a pair more related, e.g. in the pair of words 'body' and 'chest', for 'chest', the translator chose the interpretation *klatka piersiowa* (shortened to *pierś*) instead of *skrzynia*.

Unfortunately, an accurate translation of an English word into one word in Polish does not always exist. As in all

[1]The list of English words was copied from the original SimLex-999 set but new similarity values were assigned.

repetitions:	exact	reverse order
English	0	1
Italian	4	5
Russian	14	15
German	1	4

Table 1: Pair duplicates in Multilingual SimLex999 files

	T1/T2	T2/T3	T1/T3
common pairs	497	534	535
common words	1316	1374	1410

Table 2: Inter translator agreement

SimLex-999 translations only single words are used, we asked the translators to find a similarly related pair of words which is as close as possible to the original one. For example, 'groom' and 'bride' is in Polish: *pan młody* and *panna młoda*, and might be transformed to: *narzeczony* 'fiance' and *narzeczona* 'fiancée'. There were only a few such situations.

The original SimLex-999 list contains some pairs with similar words, e.g. *bad–awful, bad–terrible*. This can result in creation of two or even more identical pairs after translation. To eliminate this effect, we checked the final agreed data for repetitions, and choose alternative translations for such 22 repeated pairs. In Table 1 we show how may pairs are repeated in other Multilingual SimLex999 data.

Table 2 shows the agreement of the translators of SimLex-999 into Polish. The obtained results are worse than those reported in (Leviant and Reichart, 2015) for Russian, German and Italian, where two translators were involved in each task. The authors reported the following numbers of differently translated 1998 words: Russian 353 words, 17.7%; Italian 196 words, 9.8%; and German 396 words, 19.8%. For Polish, the best inter translator agreement is for T1 and T3 translators; they translated differently 588 words, 29.4%.

One of the reasons for discrepancies in translation are different preferences when choosing one of synonyms, e.g. in the pair *happy–chearful*, the first word was translated identically into *szczęśliwy* while the second one was either *wesoły* or *radosny*, which are near synonyms, or *pogodny*, which is a little more distant but still quite close in meaning. Sometimes, when an English pair consisted of near synonyms, one Polish word was chosen as a translation of either the first or the second word from the pair, e.g. for the pair *weird–strange* a Polish word *dziwny* was an equivalent for both *weird* and *strange*. Another source of differences is the lack of an instruction concerning perfective and imperfective verbs, e.g. *kupić* and *kupować* 'buy', which translators used inconsistently. Polish allows for diminutive forms of nouns (and even adjectives), which was another source of different translations. For example, *wuj* and *wujek* both mean 'uncle', while the first one is more official. Unfortunately, no guidelines were given for spelling. Several English words which are in common use in Polish are written differently in Polish, but the English version is more popular. For example, 'gin' can be written *dżin* or *gin*. Cultural differences meant that translators were looking for the best equivalent out of several possibilities. For instance, the differences in English and Polish education systems meant that the word 'college' was translated into three different words *uczelnia*, *uniwersytet* and *kolegż*. The best translation was the two word term *szkoła wyższa*, but, as multi-word units were excluded, we decided to use the first proposal.

3. Annotation

SimLex-999 contains information on the extent to which two words that make up a pair are similar to each other. The similarity coefficient is the average from many (approx. 50) human (Mechanical Turk) annotations. The similarity was understood here as the semantic equivalence; thus, the words that are synonyms are the most similar to each other. The annotation instruction was not very elaborate and contained two main postulates:

- words that are related are not necessary similar, e.g. *car - tire*,
- "it is perfectly reasonable to use one's intuition ..., especially when you are asked to rate word pairs that you think are not similar at all".

To retain compatibility with the original data set, we gave our annotators the same instruction. However, our annotators were linguists and computer scientists. We got three annotation from our translators who were also instructed to judge the similarity of the Polish words, but each of them only evaluated their own translation. The unified translation results – the MSimLex999_Polish dataset – were annotated by another 7 annotators who did not get the original dataset, only the Polish word pairs. These 7 annotators assigned similarity and relatedness scores to each word pair (two integers between 0 and 10). In this case, there were no formal annotation guidelines. The annotators were instructed to annotate all types of relatedness. In addition to synonymy, they also took into account hyponymy, hyperonymy, cohyponymy, antonymy, and other relations between objects or concepts that might be implied by different situations or contexts in which these objects or concepts appear in.

The final similarity and relatedness scores are average values of all annotations. To test whether this solution is plausible, we check the correlation between all the annotators' scores (pairwise). For similarity judgment the smallest Spearman's correlation coefficient (ρ) was equal to 0.52, while the highest value was equal to 0.71. To see whether the notion of the words from the source language influences the results, we compared average values for those annotators who were translators (AVR_tr) with the average obtained for all other annotators (AVG_nontr), as well as to the final average of all annotations (AVG_all). The results given in Table 3 show that differences were visible, but the final set is equally close to the judgments proposed by both groups of annotators. The average correlation between translators was a little higher (0.67) in comparison to the other group (0.63), but this might have been caused by the smaller number of translators.

	AVG_tr	AVG_nontr	AVG_all
AVG_tr	1	0.82	0.94
AVG_nontr	0.82	1	0.95
AVG_all	0.94	0.95	1

Table 3: Similarity annotation correlation between annotators who were translators and those who were not

The annotations of relatedness were a little more diverse than in the case of similarity. The highest correlation value (ρ) was equal only to 0.68, while the lowest value of the correlation was nearly the same as for the similarity – 0.53. The average correlation of the annotators was 0.59 while the average correlation with the final average ratings was 0.8.

3.1. Correlation of the Annotation with Other Languages

We compared our similarity annotation with the original SimLex-999 annotations and with those obtained by (Leviant and Reichart, 2015) while translating the data into other languages. The Spearman's coefficient for the sequences of all the pairs' similarity judgments is given in Table 4. The agreement with English data is relatively high, with greater agreement observed with respect to the original SimLex-999 annotations. The weakest agreement is observed with the German data (0.74).

dataset	correlation
SimLex-999	0.856
MSimLex999_English	0.816
MSimLex999_Russian	0.793
MSimLex999_German	0.736
MSimLex999_Italian	0.795

Table 4: Cross language similarity agreement (ρ)

The relatedness scores were compared to the association measures calculated for the SimLex-999 by University of South Florida Free Association Norms (Nelson et al., 1998) obtained directly from the SimLex-999 data. The Spearman's ρ in this case is lower (0.54). It is lower even for the correlation with the Polish similarity scores, which is equal to 0.67. The different assumptions made while assigning association scores are already visible when comparing the averages scores from these two sets. In the Free Association Norms, SimLex-999 pairs got on average association of 0.75, while in our annotations the average relatedness score is equal to 5.95. For example, the English pair *new-fresh* has the association value 1.98, while the Polish pair – 8.57, similarly the English pair *sharp-dull* has a score of 1.46, while the Polish equivalent *ostry-tępy* – 7.43.

4. Correlation of the Annotation with Polish Distributed Models

We checked the manually annotated pairs of words against several distributional models of Polish. For this purpose we prepared 16 300-dimensional models based either on forms and lemmas from the combined set of two corpora — National Corpus of Polish (Przepiórkowski, A. et al., 2012) and Polish Wikipedia. All our models were trained with the gensim tool (Řehůřek and Sojka, 2010) using two neural network architectures: Continuous Bag of Words (CBOW) and Skip-Gram (SG), and two algorithms: hierarchical softmax and negative sampling. Most of the models are described in (Mykowiecka et al., 2017a) and are available from (Mykowiecka et al., 2017b). In our experiments we also used two publicly available 100-dimensional CBOW and SG models with negative sampling trained on data from Polish Wikipedia (Rogalski and Szczepaniak, 2016a) available from (Rogalski and Szczepaniak, 2016b) – models 17 and 18 in Table 5. Table 5 contains a list of tested models.

1	f3c-hs	forms, cbow, hierarchical softmax
2	f3c-hs50	forms, cbow, hierarchical softmax, freq. above 50
4	f3c-ns1	forms, cbow, negative sampling, window 1
5	f3c-ns2	forms, cbow, negative sampling, window 2
6	f3c-ns50	forms, cbow, negative sampling, freq. above 50
7	f3s-ns	forms, skip gram, negative sampling
8	f3s-hs	forms, skip gram, hierarchical softmax
9	f3s-hs50	forms, skip gram, hierarchical softmax freq. above 50
10	f3s-ns50	forms, skip gram, negative sampling, freq. above 50
11	l3c-hs	lemmas, cbow, hierarchical softmax
12	l3c-ns	lemmas, cbow, negative sampling
13	l3c-ns1	lemmas, cbow, negative sampling, window 1
14	l3c-ns2	lemmas, cbow, negative sampling, window 2
15	l3s-ns	lemmas, skip gram, negative sampling
16	l3s-hs	lemmas, skip gram
17	plc	forms, cbow, negative sampling,
18	pls	forms, skip gram, negative sampling

Table 5: List of models. Models 1-16 are based on NKJP and Wikipedia and have 300 features. When it is not explicitly stated, the size of context window is 5. Models 17-18 are based on Wikipedia only, have 100 features and context window of the size 5.

The results of the comparison of the obtained scores with the cosine similarities of word embeddings from different models are shown in Figure 1 and 2. The results show that the correlation of the vector similarity with the manually assigned similarity scores are equally good, or even better than the correlation with the relatedness score, for the models which are based on forms and use the CBOW approach. For the form based models which use the skip gram approach and for all lemma based models, the correlation with the relatedness scores is significantly higher than with the similarity scores. Generally, lemma based models show greater correlation with both similarity and relatedness scores. This confirms the intuition that vector similarities correspond to various types of relations and not only similarity, although it was not true for all the tested models. The best model for relatedness scores is the skip gram lemma based model with negative sampling, while for the similarity, the best model is the lemma based CBOW model with negative sampling and the size of window equal

to 1. However, both models are also nearly the best in the other task. Limiting the size of the window has a limited and inconsistent influence on form based CBOW models, while it improved the correlation with the similarity scores and decreased the correlation with the relatedness scores for lemma based models. Thus, for lemma based models our conclusions are consistent with (Chiu et al., 2016).

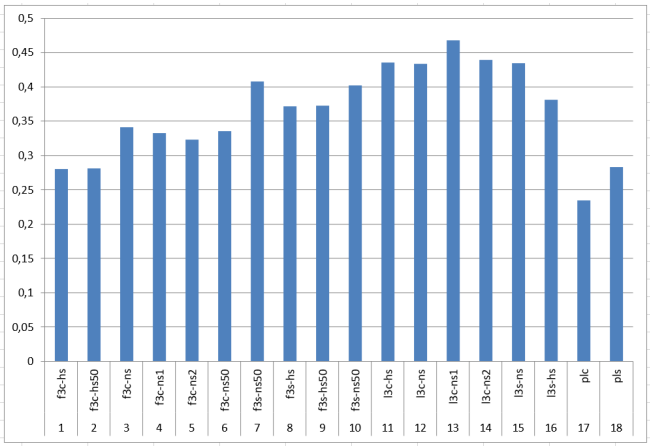

Figure 1: Correlation of MSimLex999_Polish similarities with the cosine similarity of word embeddings for different distributed models

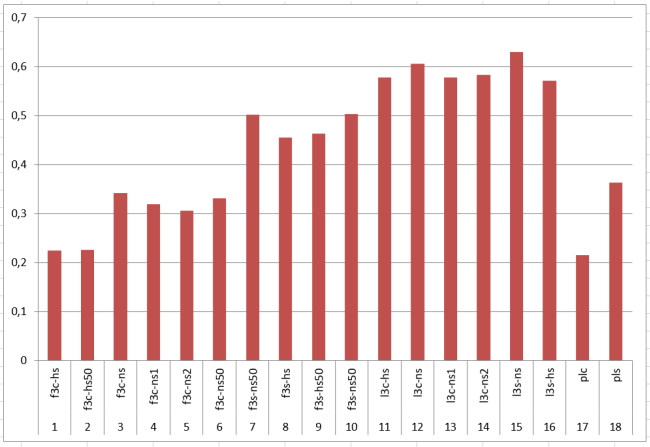

Figure 2: Correlation of MSimLex999_Polish relatedness with the cosine similarity of word embeddings for different distributed models

5. Conclusions

SimLex-999 is a resource which is frequently used as a reference set for evaluating various NLP solutions. Elaborating its Polish version may thus help in making comparisons of results of specific tasks obtained for Polish and other languages. The comparison of the manually obtained list of word similarities with the word embeddings shows that vectors obtained for a smaller data set with the smaller number of features performed worse than the larger models calculated on the bigger corpus. The best model for Polish (trained on NKJP and Wikipedia, 300 features) has the Spearman's ρ correlation with the manually annotated data equal to 0.47; while the model trained on Wikipedia, 100 features has a 0.28 correlation). The best correlation reported in (Leviant and Reichart, 2015) (a model trained on Wikipedia; 400 features) is for German — 0.34. The results obtained for the lemma based models confirmed the intuition that the vector similarity is more likely to resemble relatedness than the similarity of words, but the results for form based models are not so clear.

The relatedness judgment turned out to be more problematic. We obtained values which are different from the association values included in University of South Florida Free Association Norms and which do not show high correlation with word embeddings similarity. There may be different reasons for this and further analysis is needed.

The data are available at (Institute of Computer Science, 2017).

6. Acknowledgments

This work was supported by the Polish National Science Centre project 2014/15/B/ST6/05186 and partially as a part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

7. Bibliographical References

Chiu, B., Korhonen, A., and Pyysalo, S. (2016). Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 1–6. Association for Computational Linguistics.

Faruqui, M., Tsvetkov, Y., Rastogi, P., and Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35. Association for Computational Linguistics.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, January.

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December.

Leviant, I. and Reichart, R. (2015). Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. *CoRR*, abs/1508.00106.

Mykowiecka, A., Marciniak, M., and Rychlik, P. (2017a). Testing word embeddings for Polish. *Cognitive Studies*, 17.

Przepiórkowski, A., et al., editors. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New*

Challenges for NLP Frameworks, pages 45–50, Valletta, Malta, May. ELRA. http://is.muni.cz/publication/884893/en.

Rogalski, M. and Szczepaniak, P. S. (2016a). Word embeddings for the Polish language. In Leszek Rutkowski, et al., editors, *Artificial Intelligence and Soft Computing - 15th International Conference, ICAISC 2016, Zakopane, Poland, June 12-16, 2016, Proceedings, Part I*, volume 9692 of *Lecture Notes in Computer Science*, pages 126–135. Springer.

8. Language Resource References

Evgeniy Gabrilovich. (2017). *The WordSimilarity-353 Test Collection*. http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html.

Felix Hill. (2017). *SimLex-999*. https://www.cl.cam.ac.uk/~fh295/simlex.html.

Institute of Computer Science, Polish Academy of Sciences. (2017). *MSimLex999_Polish*. http://zil.ipipan.waw.pl/CoDeS/.

Ira Leviant and Roi Reichart. (2017). *Multilingual SimLex999 and WordSim353*. http://www.leviants.com/ira.leviant/MultilingualVSMdata.html.

Agnieszka Mykowiecka and Małgorzata Marciniak and Piotr Rychlik. (2017b). *Word embeddings for Polish*. http://dsmodels.nlp.ipipan.waw.pl/.

D. L. Nelson and C. L. McEvoy and T. A. Schreiber. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. http://www.usf.edu/FreeAssociation/.

Marek Rogalski and Piotr S. Szczepaniak. (2016b). *pl-embeddings*. http://publications.it.p.lodz.pl/2016/word_embeddings/.