Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3260–3264
Marseille, 11–16 May 2020

A Dataset of Translational Equivalents Built on the Basis of plWordNet-Princeton WordNet Synset Mapping

Ewa Rudnicka, Tomasz Naskręt
Wroclaw University of Science and Technology, Faculty of Computer Science and Management
Wybrzeże Wyspiańskiego 27, 50-370 Wroclaw
{ewa.rudnicka, tomasz.naskret}@pwr.edu.pl

Abstract

The paper presents a dataset of 11,000 Polish-English translational equivalents in the form of pairs of plWordNet and Princeton WordNet lexical units (senses) linked by three types of equivalence links: strong equivalence, regular equivalence, and weak equivalence. The resource consists of the two subsets. The first subset was built in result of manual annotation of an extended sample of Polish-English sense pairs partly randomly extracted from synsets linked by interlingual relations such as I-synononymy, I-partial synonymy and I-hyponymy and partly manually selected from the surrounding synsets in the hypernymy hierarchy. The second subset was created as a result of the manual checkup of an automatically generated lists of pairs of sense equivalents on the basis of a couple of simple, rule-based heuristics. For both subsets, the same methodology of equivalence annotation was adopted based on the verification of a set of formal, semantic-pragmatic and translational features. The constructed dataset is a novum in the wordnet domain and can facilitate the precision of bilingual NLP tasks such as automatic translation, bilingual word sense disambiguation and sentiment annotation.

Keywords: wordnet, bilingual alignment, senses

1. Introduction

Bi- and multilingual wordnets are used in a variety of tasks, such as automatic and manual translation, bilingual sentiment annotation and word sense disambiguation (Google Translate, (Bond et al., 2019)). More and more people also reach to them in addition to, or instead of, electronic dictionaries. Wordnets usually provide pairs or groups of conceptual equivalents based on mapping between synsets (sets of synonymous lexical units, also called senses). Still, such mapping may be of different granularity depending on the granularity of synsets involved (that is the number of lexical units (senses) per synset). It may include 1-1 mappings (for singleton synsets), 1-to-many mappings (for singleton and multi-unit synsets) and many-to-many mappings (for multi-unit synsets). Intuitively, for all the tasks mentioned, we predict that precision should rise directly proportional to the precision of bilingual links. The most precise links obviously hold between singleton synsets. In fact, this moves mapping from the level of synsets to the level of individual sense pairs. However, such mappings are rather rare in the wordnet domain, because its automatization is difficult, while manual work extremely costly and time-consuming. The notable exceptions are the works of (Copestake et al., 1994) and our own (Rudnicka et al., 2017b), (Rudnicka et al., 2017a), (Rudnicka et al., 2018).

In our most recent paper (Rudnicka et al., 2019), we have refined our earlier feature-based typology of equivalence types together with the method of their application and have tested it on a sample of lexical unit pairs. In this paper we will show the results of applying this method on a larger scale. We are going to present a dataset of plWordNet-Princeton WordNet sense mappings constructed in result of two tasks: manual mapping of an extended sample of sense pairs and manual evaluation of an automatically generated set of sense mappings. Both types of mappings are based on the results of a large scale synset mapping between plWordNet and Princeton WordNet (Rudnicka et al., 2012). They take as its input sense pairs extracted from Polish and English synsets linked mainly by interlingual synonymy relation (construed as simple equivalence) and, in manual mapping, also by interlingual partial synonymy and hyponymy relations.

The automatic linking was generated on the basis of manual synset mapping, automatic extraction of knowledge from a cascade of knowledge sources such as selected Polish-English dictionaries, Wiktionary and Wikipedia, and manual annotation relying on checking equivalence features. Thus, the linking involved two key stages. The first one was the automatic generation of Polish-English sense pairs from interlingual synset relations supported by bilingual cascade dictionary data (describe cascade dictionary). The second one consisted in manual annotation of the automatically generated sense pairs (by equivalence types) on the basis of predefined equivalence features. In this way, the results of automatic sense alignment gained human verification.

The paper will be structured as follows. First, we will describe a method of manual synset mapping and a method of automatic generation of sense pairs. Next, we will present a method of manual annotation of automatically produced sense pairs. Then, the properties of the constructed data set will be illustrated by different types of statistics. The paper will close with conclusions and directions for future work.

2. From Synset to Sense Mapping

plWordNet is one of the few wordnets built via a corpus-based variant of the so called merge wordnet-building method (Piasecki et al., 2009) and only later manually mapped to Princeton WordNet on synset level (Rudnicka et al., 2012). The merge method consists in building a wordnet from scratch on the basis of data collected from available lexicographic resources and corpora and verified by lexicographers. Other wordnets constructed by means of the merge method are GermaNet(Kunze and Lemnitzer, 2002), DanNet(Pedersen et al., 2019), and Czech Word-

Net(Pala and Smř, 2004). Such method of lexico-semantic resource construction has its pros and cons. It allows to produce a wordnet not biased towards English, but it lacks on providing simple equivalences between its synsets. Still, it offers a more reliable picture of a particular language lexicon and its relation to another language lexicon.

Currently, synset mapping of plWordNet to Princeton WordNet is almost complete for nouns, adjectives and adverbs, while the mapping of verbs is still in progress (see the statistics http://plwordnet.pwr.edu.pl/wordnet/stats). It involves a rich network of interlingual relations such as interlingual synonymy, partial synonymy, inter-register synonymy, hypo- and hypernymy, mero- and holonymy, cross-categorial synonymy, type/instance and a set of specific verb relations. They allow to capture different types of correspondence between English and Polish. The mapping has been carried out by a team of supervised lexicographers, so its quality is reasonably high[1].

In the course of synset mapping we noted a potential for creating a more fine-grained sense mapping, especially on the basis of interlingual synonymy relation, (Rudnicka et al., 2017b), (Rudnicka et al., 2017a). Building on that observation, we developed a method of manual sense mapping, (Rudnicka et al., 2017b), (Rudnicka et al., 2018), (Rudnicka et al., 2019). It capitalises on the results of synset mapping to the extent that in the initial step candidate pairs of Polish-English lexical units (senses) are extracted from synset links. The method relies on a manual checkup of pairs of Polish-English lexical units with respect to formal, semantic-pragmatic and translational features such as number, gender, countability, sense (denotation), lexicalisation of concepts, register, collocations, context, dictionary listing, position of dictionary equivalent, and parallel corpus hits. On the basis of the chosen values of respective features, lexicographers determine the strength of equivalence holding between Polish and English lexical units within a given pair. Three types of links are distinguished: strong equivalence, regular equivalence and weak equivalence. For strong and regular equivalence, the values of equivalence features need to be identical or very close. For weak equivalence, the requirements are less rigid, but still the possibility of use as at least a component of a descriptive equivalent is necessary (in the sense of (Svensen, 2009)). When the above criteria are not fulfilled an automatic link is tagged as 'mismatch'. When there is no enough data available to make a well-grounded decision, the tag 'no decision' is assigned.

The proposed method was tested on a sample of Polish-English lexical unit pairs extracted from plWordNet and Princeton WordNet synsets linked by interlingual relations such as I-synonymy, I-partial synonymy and I-hyponymy. Such move was motivated by a prediction already signalled in our earlier work (Rudnicka et al., 2018), (Rudnicka et al., 2019) that partial synonymy and hyponymy relations may also be a source of translational equivalents, esp. of regular or weak type. The samples were randomised with respect to synset granularity and type of link (1-1, 1-many, many-many) and type of interlingual relation. For the latter, the frequency of occurrence in the overall noun synset mapping was also taken into account. For the first experiment, described in (Rudnicka et al., 2019), three samples, 120 pairs each, were annotated by two lexicographers. The results of their annotation were presented in the form of confusion matrix (Rudnicka et al., 2019).

Later, the works on the annotation of the remaining samples were continued. Lexicographers also extended the mapping to pairs of lexical units to synsets located in the direct neighbourhood (within the hypernymy hierarchy) of the synsets originally selected for sample sense mapping. In this way, they managed to create a data set of 6 690 equivalence links between Polish and English lexical units. The counts of specific equivalence relation types are given in Table 1, while their respective percentages are illustrated in Figure 1.

Relation type	Count
strong equivalence	5466
regular equivalence	908
weak equivalence	316
TOTAL	6690

Table 1: Manually created dataset in numbers

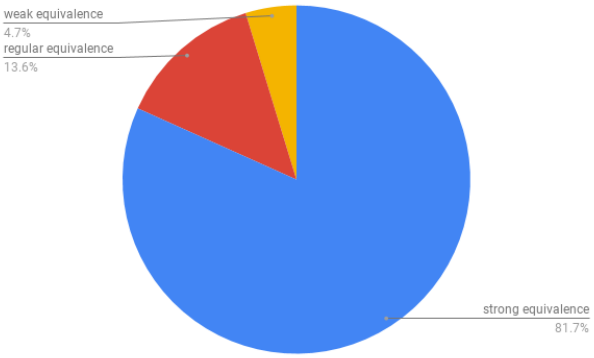

Figure 1: Data distribution in manually created dataset

We observe that the majority of links form strong equivalence links (81,7 percent). Regular equivalence links are noted for 13,6 percent of cases, while weak equivalence links were assigned only to 4,7 percent of links. Such distribution of equivalence types is (at least partially) due to a method of data set construction. In the original sample set, lexical unit (sense) pairs extracted from many-to-many synset relations constituted only one fourth of a sample, while the remaining pairs were extracted from singleton synset links, most prone to yield strong equivalence, and 1-to-many and many-to-1 synset links, also likely to contribute strong equivalence links. Later, when working on the extension of the data set, lexicographers assumed 'pick the low hanging fruit' method while analysing and linking pairs that seemed likely to be good equivalents at the very first glance.

In Figure 3, we present the distribution of strong equivalence relation across semantic domains (lexicographer files) of the Polish and English lexical units linked by this relation. We

[1] Our Polish-English wordnet data is used by Google Translate, and a couple of online dictionaries.

observe a wide variety of domains represented in the constructed data set. Clearly, the highest frequency of links holds for pairs whose units share the semantic domain, e.g. artefact (19 percent), thinking (10,3 percent), communication (9,6 percent), person 10,2 percent), food (5,9 percent), and quantity (5,2 percent).

3. A Method of Automatic Sense Alignment

Inspired by our earlier work (Rudnicka et al., 2017b), we tested the possibility of automatic extraction of pairs of translational, sense-level equivalents from Polish-English pairs of synsets linked by inter-lingual synonymy relation. For these purposes, we used a couple of simple heuristics, in some cases supported by the data from a cascade Polish-English dictionary and from Wikipedia.

In the first step, we generated all possible Polish-English pairs of lexical units from synsets linked by interlingual synonymy relation. Next, from the whole set of these Polish-English pairs we filtered all pairs of Polish-English lexical units that shared a common lemma. The obtained 'identical lemma' subset consisted of 3687 pairs. In the next step, we took advantage of the existing mapping between the subset of plWordNet lexical units and directly corresponding Wikipedia articles. We selected semantic domains (lexicographer files) such as substance and quantity where the mapping was most likely to hold. We filtered all pairs of Polish-English lexical units where the lemma of the English unit was identical to the lemma of the title of the English Wikipedia article equivalent to the Polish Wikipedia article mapped to the Polish unit. We obtained 1326 pairs for substance and 155 pairs for quantity.

For the remaining steps, we decided to use the so called cascade dictionary combined from a variety of electronic Polish-English dictionaries (made available to us under open licence) as well as Wiktionary and Wikipedia. From our initial set of Polish-English lexical units, we extracted such pairs in which the lemma of the English unit was the only translation given by the dictionary. The obtained subset amounted to 8057 pairs. The last idea was to use 'femininity' relation from plWordNet which links feminine gender nouns to their masculine derivational bases. It is not always the case that a derivative is the feminine form of the same word. Therefore, we reached to the cascade dictionary again and extracted only those pairs in which the neutral gender English noun is given as an equivalent of a feminine gender Polish noun. The obtained subset counted only 73 pairs. The set of pairs generated in result of applying all heuristics amounted to 13 298 pairs of Polish-English lexical units. We deleted pairs that appeared more than once in the set (since they were generated in result of applying more than one heuristics) and that gave us 9 146 pairs. In last step, we reduced the set to the set to unique links between specific pairs of lexical units, because many links were represented twice: going from a Polish to an English unit and the other way round. The final set counted 5 060 pairs.

The next step of the resource construction was manual verification of the results of automatic sense alignment. For these purposes, we employed the same method we used for manual sense alignment already described in Section 2. The detailed distribution of equivalence types is given in Table 3 and its percentages are shown in Figure 2. We observe a very high agreement of lexicographers' choices with the results of automated linking. In 89,5 percent of cases they agreed that a postulated link is in fact a strong equivalence link. They opted for regular equivalence in 4,6 percent of cases, while weak equivalence was assigned in only 0,7 percent of cases (34 pairs). Mismatch cases constitute only 4,7 percent of cases, while no decision concerned only 27 pairs. The constructed set of manually checked mappings consists of 5 064 pairs.

Heuristic type	Count
femininity	73
identity of the Polish and English lemma	3687
single translation in the PL-Eng dictionary	8057
Wikipedia substance	1326
Wikipedia quantity	155
TOTAL	13298

Table 2: Heuristics types and the counts of the produced pairs

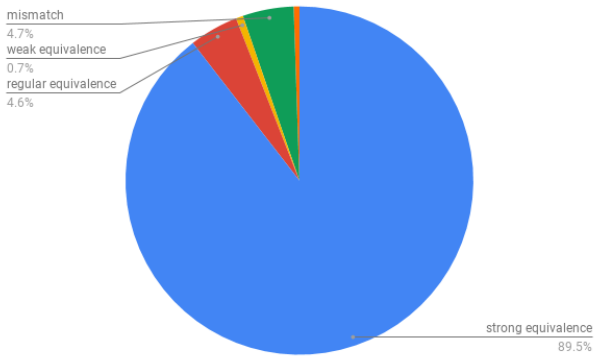

Figure 2: Automatically created dataset in numbers

Such results can be explained by the type of heuristics that were used in generating automatic pairs, see Table 2. The heuristics that contributed the larger number of pairs checked if the lemmas of the Polish and English lexical units are its only translation in a bilingual dictionary. Clearly, if they are such, they are very likely to be strong equivalents. The second in the number of pairs heuristic relied on the identity of the Polish and English lemma. This one could have yielded the so called false friends, if it was not for the fact that the pairs were drawn only from synsets linked by interlingual synonymy. Heuristics based on the mapping to the interlinked Wikipedia articles were also likely to contribute good equivalents.

Similarly, to our fully manually created subset, the distribution of strong equivalence relation across semantic domains is very varied, as illustrated in Figure 4. Again, the highest frequency holds for pairs for lexical units that share a semantic domain such as substance (24,7 percent), artefact (8 percent), person (7,1 percent), quantity (5,2 percent).

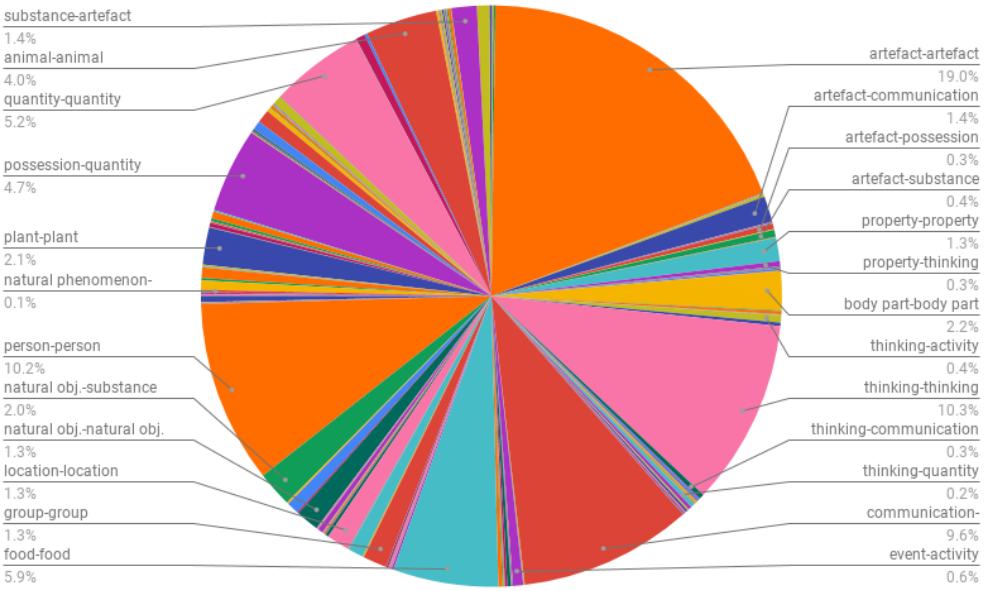

Figure 3: Strong equivalence distribution across domains in manually collected data subset.

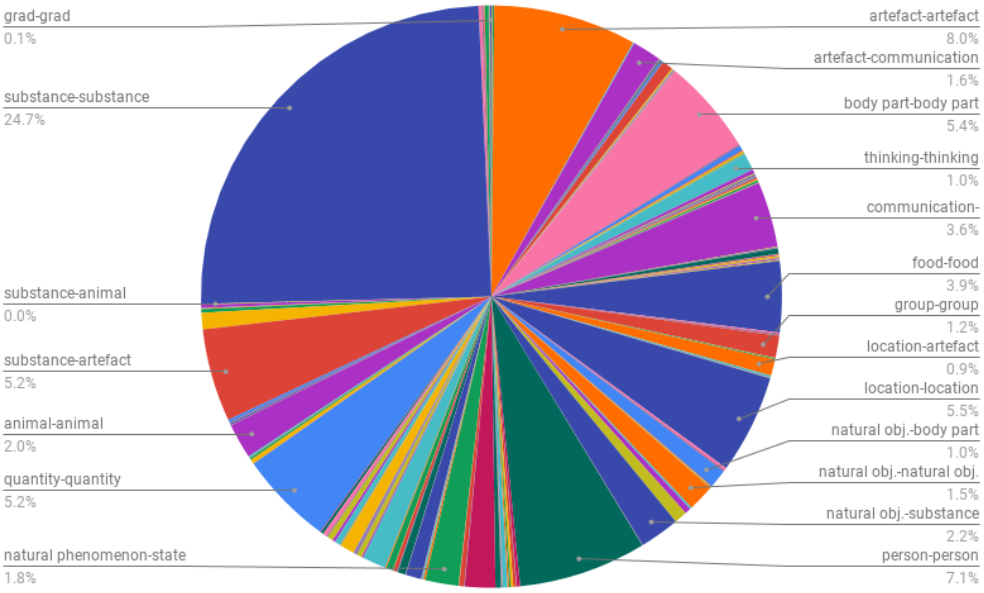

Figure 4: Strong equivalence distribution across domains in an automatically created data subset.

Relation type	Count
strong equivalence	4532
regular equivalence	231
weak equivalence	34
mismatch	240
no decision	27
TOTAL	5064

Table 3: Data distribution in automatically created dataset

4. Properties of the Produced Dataset

The final dataset that was built in result of applying manual annotation and an automatically enhanced manual annotation counts 11 487 equivalence relations established between pairs of Polish-English lexical units, as shown in Table 4. As its two component subsets, it shows the highest frequency for the strong equivalence relation, which is a welcome results in view of the aim of its construction that is creating a dataset that would possibly foster the precision of bilingual NLP tasks.

The described dataset constitutes an integral part of the plWordNet (Pol. Słowosieć) database which is available for

Relation type	Count
strong equivalence	9998
regular equivalence	1139
weak equivalence	350
TOTAL	11487

Table 4: Properties of the final dataset

download [2]

5. Conclusion

As a result of works described in the paper, we have managed to build a 11 000 data set of Polish-English manual sense links having the power of translational equivalents. To our knowledge no similar data set of such size and link precision exists for any other wordnet. Since bi- and multilingual wordnets are more and more often used as dictionaries and are also one of the basic resources for automatic translators and word sense disambiguation tools, the existence of sense level links gains more and more necessity.

6. Acknowledgements

Co-financed by the Polish Ministry of Education and Science, CLARIN-PL Project..

7. Bibliographical References

Bond, F., Janz, A., and Piasecki, M. (2019). A comparison of sense-level sentiment scores. In Christiane Fellbaum, et al., editors, Proceedings of the Tenth Global Wordnet Conference, pages 363–372.

Copestake, A., Briscoe, T., Vossen, P., Ageno, A., Castellón, I., Ribas, F., Rigau, G., Rodríguez, H., and Samiotou, A. (1994). Acquisition of lexical translation relations from mrds. *Machine Translation*, 9:183–219, 09.

Kunze, C. and Lemnitzer, L. (2002). GermaNet – representation, visualization, application. In Proc. LREC 2002, main conference, volume V, pages 1485–1491.

Pala, K. and Smř, P. (2004). Building czech wordnet. *ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY Volume*, 7:79–88, 01.

Pedersen, B., Nimb, S., Asmussen, J., Sørensen, N., Trap-Jensen, L., and Lorentzen, H. (2019). Dannet - a wordnet project for danish. 12.

Piasecki, M., Szpakowicz, S., and Broda, B. (2009). A Wordnet from the Ground Up. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

Rudnicka, E., Maziarz, M., Piasecki, M., and Szpakowicz, S. (2012). A Strategy of Mapping Polish WordNet onto Princeton WordNet. In Proc. COLING 2012, posters, pages 1039–1048.

Rudnicka, E., Bond, F., Grabowski, , Piasecki, M., and Piotrowski, T. (2017a). Towards equivalence links between senses in plWordNet and princeton WordNet. *Lodz Papers in Pragmatics*, 13(1):3–24.

Rudnicka, E., Piasecki, M., Piotrowski, T., Łukasz Grabowski, and Bond., F. (2017b). Mapping wordnets from the perspective of inter-lingual equivalence. *Cognitive Studies / Études cognitives*, 17. in print.

Rudnicka, E., Bond, F., Grabowski, L., Piasecki, M., and Piotrowski, T. (2018). Lexical perspective on wordnet to wordnet mapping. In Proceedings of the 9th Global Wordnet Conference, 01.

Rudnicka, E., Bond, F., Grabowski, Ł., Piotrowski, T., and Piasecki, M. (2019). Sense Equivalence in plWordNet to Princeton WordNet Mapping. 03.

Svensen, B. (2009). A Handbook of Lexicography. The Theory and Practice of Dictionary-Making. Cambridge University Press., Cambridge.

[2] http://ws.clarin-pl.eu/public/wordnet-work.LATEST.sql.gz