Challenges of Adjective Mapping between plWordNet and Princeton WordNet

[1]Ewa Rudnicka, [2]Wojciech Witkowski, [3]Katarzyna Podlaska
[1,3]Wrocław University of Technology, [2]University of Wrocław
Poland
email: ewa.rudnicka@pwr.edu.pl, wojciech.witkowski@uwr.edu.pl, k.podlaska@hotmail.com

Abstract

The paper presents the strategy and results of mapping adjective synsets between plWordNet (the wordnet of Polish, cf. Piasecki et al. 2009, Maziarz et al. 2013) and Princeton WordNet (cf. Fellbaum 1998). The main challenge of this enterprise has been very different synset relation structures in the two networks: horizontal, dumbbell-model based in PWN and vertical, hyponymy-based in plWN. Moreover, the two wordnets display differences in the grouping of adjectives into semantic domains and in the size of the adjective category. The handle the above contrasts, a series of automatic prompt algorithms and a manual mapping procedure relying on corresponding synset and lexical unit relations as well as on inter-lingual relations between noun synsets were proposed in the pilot stage of mapping (Rudnicka et al. 2015). In the paper we discuss the final results of the mapping process as well as explain example mapping choices. Suggestions for further development of mapping are also given.

Keywords: wordnets, inter-lingual mapping, adjective relation structure

1. Introduction

The goal of this paper is to present solutions developed for the purposes of mapping two different relation structures describing adjectives in Princeton WordNet (henceforth, PWN, cf. Fellbaum 1998) and in plWordNet (Polish wordnet, henceforth, plWN, cf. Maziarz et al. 2013). plWordNet is one of the few world wordnets built fairly independently of PWN with the help of a unique method of lexico-semantic relations extraction from large text corpora (Piasecki et al. 2009). Nevertheless, the actual construction process is manual - a supervised team of lexicographers verifies automatic hints in lexicographic resources and only then introduces them into a database. Thus, it belongs in with the so called *merge* approach (cf. Vossen et al. 2002). It allows for a more truthful language description, but leads to differences in lexical coverage and relational structures between wordnets. This is clearly the case of plWN and PWN adjective domain, which has vertical, hyponymy-based structure in plWN (akin to that of nouns and verbs) (cf. Maziarz et al. 2012), and a horizontal, *dumbbell* model-based structure in PWN (cf. Miller 1998, Sheinman et al. 2013). Moreover, plWN has a slightly more fine-grained set of semantic domains comprising qualitative, relational and material adjectives, while PWN distinguishes only relational adjectives from the general adjective category. Another key issue in the process of adjective mapping is wordnet size. At the beginning the sizes of plWN and PWN adjective domains were comparable[1]. However, the process of mapping has been carried out parallel to the process of the extension of adjective category in plWN, and at the final stage of mapping the number of adjective synsets in plWN outgrew that of PWN twice[2].

In view of the above mentioned contrasts, the design of the mapping strategy for plWN and PWN adjectives had been a real challenge. We started with a detailed analysis of both synset and lexical relation structures with an eye to any common points between the two wordnets. Wordnet mapping is carried out at the level of synsets (cf. the EuroWordNet project, Vossen 2002; OpenMultilingual WordNet, Bond et al. 2013), but here lexical units relations looked much more promising. We designed a series of rule-based, automatic prompt algorithms capitalising on corresponding synset and lexical unit relations in the two wordnets and on the already existing inter-lingual noun synset relations (cf. Rudnicka et al. 2012, Rudnicka et al. 2015). The latter was possible because many of the adjective relations are relations to nouns (Maziarz et al. 2012). The rules were accompanied by lemma filtering of the achieved synset pairs by a large Polish-English cascade dictionary, similarly as in the process of generating automatic prompts for nouns (cf. Kędzia et al. 2013). We also took advantage of noun mapping experience in drawing a procedure for manual mapping and an inventory of inter-lingual relations. Main relations stayed the same including synonymy, hyponymy and partial synonymy, but their definitions had to be adjusted to the specificity of adjective category. Also, new relations had to be added, especially varieties of cross-categorial synonymy to nouns.

The paper is structured as follows. Section 2 offers a comparative analysis of adjective relation structure in plWN and in PWN, Section 3 describes our proposal of

1 This is based on the data from plWN 2.1 version, downloadable from http://nlp.pwr.wroc.pl/plwordnet/download/?lang=pl

2 All the counts given throughout this paper are taken from the official plWordNet website: http://plwordnet.pwr.wroc.pl/wordnet/stats.

the manual mapping procedure and automatic prompt algorithms, Section 4 presents the discussion of the achieved results. The paper closes with conclusions and suggestions for further research given in Section 5.

2. Adjectives in plWordNet and in Princeton WordNet

The very idea of a wordnet-type dictionary has its origins in the psycholinguistic research of 60-ties of the XX-tieth century (cf. Collins and Quillian 1969). The structure of the original Princeton WordNet was aimed to map the structure of human lexical memory (cf. Fellbaum 1998): nouns and verbs were organised into hierarchical, hyponymy-based structures, adjectives and adverbs into antonymy-based opposition structures (cf. Deese 1964). Antonymy is a relation between specific word forms in specific senses, hence it is established on the level of *lexical units* (lemma sense pairs, the smallest wordnet building blocks). Hyponymy is a relation between concepts, hence it is established on the level of s*ynsets* (sets of synonymous lexical units, main wordnet building blocks). To link adjective synsets, a special *Similar to* relation was proposed (cf. Miller 1998). It groups them into the so called *dumbbells*. These are sets of closely semantically related adjective synsets organised around a central adjective synset whose lexical units are linked by *Antonymy* relation to their semantically opposite lexical units of a central adjective synset of another dumbbell (cf. Miller 1998, Sheinman et al. 2013). Below we illustrate the dumbbell model with printscreens from the WordNetLoom editing tool. Synset relation structures for the adjecitives *small* and *large* are given, together with Antonymy relation between their lexical units.

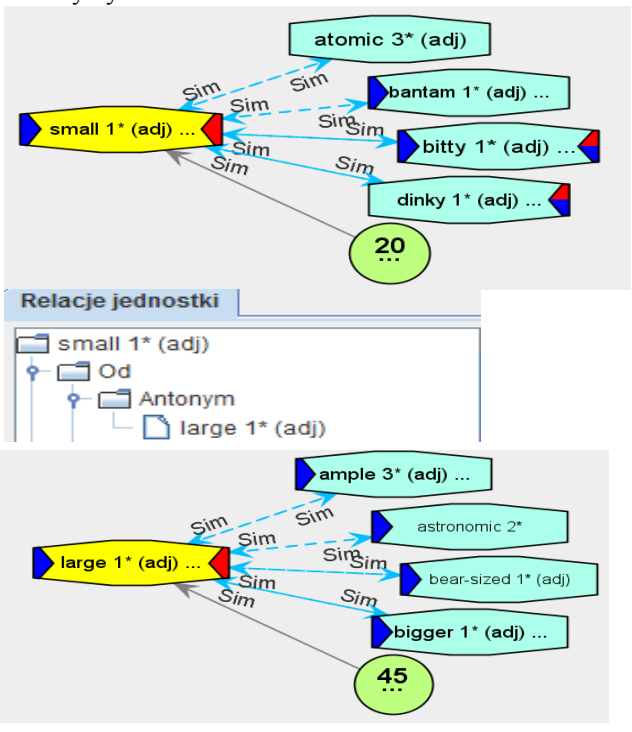

Figure 1. Application screenshot representing adjective relation network in PWN

Notwithstanding the psycho-linguistic reality of the dumbbell model, it is criticised for hindering natural language processing tasks such as semantic similarity measure critical for word sense disambiguation (cf. Sheinman et al. 2013). Crucially, these require hierarchical, hyponymy based-structures.

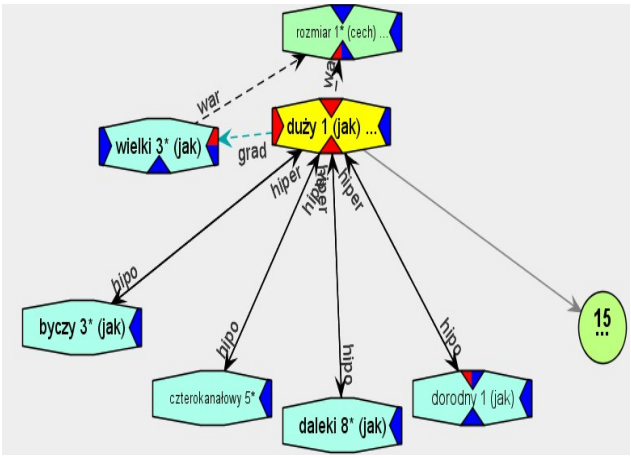

Figure 2. Application screenshot showing the structure of adjective synset relation network in plWN

Such structures have been developed for adjective synsets by the constructors of plWordNet. Apart from *Antonymy* relation between adjective lexical units, they defined *Hyponymy* relation between adjective synsets (cf. Maziarz et al 2012). Thus, adjective domain in plWordNet has vertical structure, akin to that of nouns and verbs. Again, it is illustrated in the screenshot from the WordNetLoom editing tool given in Figure 2 below. It shows the structure of synset relations for the Polish adjective *duży* – 'large'.

The graph structure given in Fig. 2 shows the set of hyponyms for the adjective *duży* - 'large' represented by vertical black lines with the index 'hipo'. Still, apart from *Hyponymy* relations, also other relations are visible on the graph: 'war' standing for *Value of the attribute* and 'grad' standing for *Gradability*. The whole list of plWN and PWN synset relations and their counts is given in Table 1:

Synset relation counts		
Relation	*plWN*	*PWN*
(Value of the)/ Attribute	9658	639
Modifier	2108	------
Hyponymy	18225	------
Gradability	991	------
Near-synonymy	1308	------
Similar to	------	21434
Member of this domain	------	1418

Table 1. plWN and PWN adjective synset relation counts

The data in Table 1 clearly show that the set of adjective synset relations in plWN and in PWN is very different, which signals future problems in mapping between the two networks. The only directly corresponding relation is

Value of the)/Attribute, but its count in PWN is very low. Apart from the main *Similar to* relation, PWN has Member of this domain relation with the subtypes *Topic, Region, and Usage.* Such type of information is rendered in plWN by register labels attached on the level of lexical units. Apart from the main Hyponymy relation, plWN also has more specific relation such as *Modifier*, *Near-synonymy*, and *Gradability*, which pertain to better differentiation of synset meanings (concepts).

Luckily, lexical unit relations in plWN and in PWN bear much more resemblance, as shown in Table 2 below:

Lexical unit relation counts		
Relation	*plWN*	*PWN*
Antonymy	5318	4024
Cross-categorial synonymy/Pertainym	15139	3293
Derivativity/Derivationally related form	11653	14317
Similarity	1959	------
Characterising	4974	------

Table 2. plWN and PWN adjective lexical unit relation counts

Apart from Antonymy relation, plWN and PWN have two more directly corresponding relations: *Derivativity* and *Derivationally related form*, and *Cross-categorial synonymy* and *Pertainym*. The tracked correspondences have been utilised in the development of automatic prompt algorithms discussed in more detail in Section 3.

Another area of contrast between plWN and PWN constitute semantic domains adjectives are grouped into. In plWN, adjectives are divided into relational, quality-denoting and material-denoting, while in PWN only relational adjectives are singled out, the remaining ones are not further classified and appear under a general 'adj' heading.

Domain counts		
Domain	*plWN*	*PWN*
[jak] quality-denoting	23644	-------
[rel] - relational	14843	3665
[adj] - adjective	-------	14460
[mat] - material-denoting	1118	-------

Table 3. plWN and PWN adjective domain counts (in lexical units) compared

The domains provide information on the semantic content of adjectives and as such the differences in their number and counts will also need to be reflected by appropriate choices when establishing inter-lingual relations. At least the criteria for distinguishing relational adjectives are similar in plWN and in PWN. In plWN, they need to be linked to nouns by *Cross-categorial synonymy* relation, while in PWN by *Pertainym* relation. The two relations are largely corresponding in terms of their semantic import.

The last key issue in the process of adjective mapping has become wordnet size. At the beginning of the mapping process the sizes of plWN and PWN adjective category domains were comparable. However, the process of mapping has been carried out parallel to the process of the extension of adjective category in plWN. The counts of the latest official plWordNet 2.3 version are presented in Table 4 below:

Basic counts		
	plWN	*PWN*
no. of lemma	26961	21808
no. of lexical units	45514	30072
no. of synsets	38668	18185

Table 4. plWN and PWN adjective basic counts compared

As shown in Table 4, plWN outgrows PWN in the number of all basic building blocks: lemmas, lexical units and synsets. The contrast is the sharpest (i) in the case of synset counts – their number is over two times higher in plWN than in PWN and (ii) in the ratio of lexical units per synset (1.17 LU/synset in plWN; 1.65 LU/synset in PWN). This already signals potential problems in the mapping process, especially difficulties in establishing (full) *Inter-lingual synonymy* relation links between plWN and PWN synsets.

3. Mapping strategy

In designing the strategy for mapping adjective synsets between plWN and PWN, we focused on relations common or similar in the two networks and used them as a starting point for the first stage of mapping (cf. Rudnicka et al. 2015). Two types of algorithms generating automatic prompts were developed. The first one relied on synset relations, exclusively, such as *Attribute* and *Value of the attribute a*nd *Similar to* and *Hyponymy* and *Gradability.* The second one capitalised on both synset and lexical unit relations taking in addition *Derivationally related form* and *Derivativity*, *Pertainym* and *Cross-categorial synonymy*. Apart from adjective relations, both algorithms took advantage of the existing network of inter-lingual relations between noun synsets, because some of adjective relations are relations to nouns. Finally, lemmas of the generated candidate pairs were filtered by a large cascade dictionary. The results of sample implementation of the algorithms were next confronted with the results of independent manual mapping. Tests showed higher effectiveness of the 'mixed' type of algorithm (figures from TSD).

Notwithstanding the usability of the proposed automatic prompt algorithms, it was necessary to design the manual mapping procedure and define a set of inter-lingual relations. In its main assumptions, it follows the general mapping procedure proposed for nouns (cf. Rudnicka et al. 2012). The procedure consists of three

main stages: recognising the sense of a source synset, searching for a target language synset and establishing a relevant inter-lingual relation. The set of inter-lingual relations also largely corresponds to a set of inter-lingual relations defined for the purposes of noun mapping and it includes *Synonymy, Partial synonymy, Inter-register synonymy, Hyponymy, Hypernymy* and, in addition, *Cross-categorial synonymy* to nouns. The latter relation is used in the cases of very general *I-hyponymy* links for more detailed specification of the sense of a source synset. Three subtypes are distinguished: *made of* used for adjectives describing a material denoted by a noun, *resembling* used for adjectives naming a physical property denoted by noun and *related to* used for adjectives describing a non-physical property denoted by a noun.

4. Mapping results and discussion

In this section, we discuss the results of mapping adjective synsets in plWN 2.3. Tables 5 and 6 provide the counts of inter-lingual relations and their distribution across wordnet domains:

I-relation counts	
Relation	*Instances*
I-synonymy	3549
I-partial synonymy	1397
I-inter-register synonymy	47
I-hyponymy	17654
I-hypernymy	63
I-cross-categorial synonymy	13792
TOTAL	36455

Table 5. Adjective inter-lingual relation counts

Cross-domain inter-lingual relation counts		
plWN / PWN	*[adj]*	*[rel]*
[jak] - quality-denoting	11035	2271
[mat] - material-denoting	550	479
[rel] - relational	3875	10172

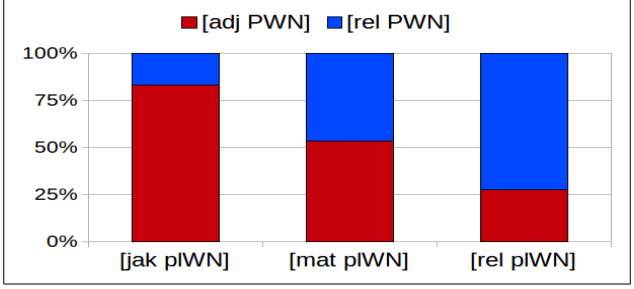

Table 6. Adjective cross-domain interlingual relation counts

Bearing in mind size differences of the adjective category in plWN and PWN (see Section 2), it comes as no surprise that the most frequent inter-lingual relation is *I-hyponymy*, with 17654 links (48%). It is followed by *I-cross-categorial synonymy*, with 13792 links (38%). Still, it must be remembered that *I-cross-categorial synonymy* is established only as a 'follow-up' of *I-hyponymy* when its links are very general. The majority of these links are cases of Polish adjectives derived from nouns by productive morphological rules and having no direct adjective counterparts in English (78% of the mapped adjectives with *I-hyponymy* relation). Thus, if we subtract the number of those cases from the total number of *I-hyponymy* links, we obtain 3862 'true' *I-hyponymy* links and 22663 plWN adjective synsets linked to PWN adjective synsets by non-cross-categorial inter-lingual relations. Also, the contrast in *I-hyponymy* and *I-synonymy* counts becomes then much less sharp. It is only 3862 against 3549 links. *I-synonymy* provides the most specific type of link, hence it is always the most desired relation from the perspective of mapping. The last notable inter-lingual relation is partial synonymy with 1397 links.

Productive affixal adjective derivation in Polish is not the only source of mapping problems. Another issue is domain mismatch. plWN and PWN semantic domains only partially overlap (see Section 2). This motivates the necessity of establishing inter-lingual relations between plWN and PWN adjective synsets belonging to different domains. In Table 6, we present the statistics of cross-domain mappings. The most interesting piece of data is the number of links established between plWN quality-denoting adjectives and PWN relational adjectives, which is 2271. One would expect plWN quality-denoting adjectives to be mapped rather on PWN adjectives from the general 'adj' domain. The most troublesome cases are those in which plWN quality-denoting adjectives and PWN relational adjectives are to be linked by means of (full) *I-synonymy* due to the fact that establishing any other inter-lingual relation would falsify the semantic correspondence between lexical units forming the synsets in question. In such cases, lexicographers choose to violate the requirements posed by wordnet design, i.e. not using *I-synonymy* to link adjective synsets of different domains, and establish *I-synonymy* relation between the synsets. To illustrate this, consider the synsets {postkomunistyczny 2 (jak)} - 'postcommunist' (quality-denoting) and {post-communist 2 (rel)}:

Figure 3. Application screenshot of the relation network of the Polish synset {postkomunistyczny 2 (jak)}

Both {postkomunistyczny 2 (jak)} 'occurring after communism' and {post-communist 2 (rel)} 'no longer communist; subsequent to being communistic' refer to

events that happen after communism. Semantically, they can be thought of as equivalents. Furthermore, they have corresponding positions in the synset networks in respectful wordnets. Yet, their qualifiers (jak) - 'quality-denoting' and (rel) - 'relational' indicate that in plWN the adjective *postkomunistyczny* denotes the quality of an object it modifies, i.e. occurring after communism is an inherent quality of a given object, whereas in PWN *post-communist* denotes relation to communism, i.e. the object modified might as well occur during communism and continue its occurrence after it. Naturally, the decision to establish full *I-synonymy* between the two synsets irrespective of domain differences raises the question whether the emphasis should be put on the mapping of senses or the mapping of structures. A question which we leave unanswered for the time being.

5. Conclusion

Mapping between two independently created networks is always a challenge. In plWN-PWN adjective mapping, we had to deal with relation structure, semantic domain and size differences between the two wordnets. To handle them, an advanced mapping strategy was proposed subsuming a three stage mapping procedure, a set of inter-lingual relations and automatic prompt algorithms. Both manual mapping procedure and automatic prompt algorithms capitalise on pairs of relations that are corresponding between the two wordnets, such as, for instance, *Value of (the Attribute)*, *Derivativity/Derivationally related form* and *Cross-paradigm synonymy/Pertainym*. The algorithms also take advantage of inter-lingual noun mapping between plWN and PWN, since some of intra-wordnet adjective relations are relations to nouns.

The most frequently established inter-lingual relation is *I-hyponymy*, yet the vast majority of these links result from morphological differences between English and Polish, namely very productive affixal derivation of Polish adjectives from nouns. Many of these adjectives do not have direct equivalents in English. To make their semantic import more specific we have introduced *I-cross-categorial synonymy* relation to English nouns and its links comprise about three fourths of *I-hyponymy* links. The remaining number of *I-hyponymy* links is comparable to the number of *I-synonymy* links. High frequency of *I-cross-categorial synonymy* appears inevitable and is dependent on the number of noun derived adjectives in plWN and the number of noun synset pairs that exist in linked plWN and PWN. The fact that *I-cross-categorial synonymy* works in correlation with *I-hyponymy* will undoubtedly increase the number of *I-hyponymy* links characterised by a limited information input as far as semantic relations between inter-lingually linked adjectives are concerned. With respect to future works, *I-cross-categorial synonymy* could gain more fine-grained distinctions that will allow lexicographers to establish more precise meaning correspondences between Polish derived adjectives and English nouns that correspond to Polish nouns which are bases for the derived adjectives, e.g. *I-cross-categorial synonymy* of the type *Related to* could be further divided into relations coding narrower semantic correspondences, for instance *(Related_to)_Location* for adjectives derived from place names;

{Warsaw}---I-cross-categorial------
synonymy_(Related_to)_Location----
{warszawski (Warsaw(Adj))}
← {Warszawa (Warsaw(N))}.

Different semantic domains of adjective synsets to be linked have turned out secondary to their semantic closeness. The choice of meaning correspondence over even the grammatical category mismatch is nothing uncommon as far as multi-lingual wordnets are concerned. A parallel approach is visible for instance in EuroWordNet (see Vossen 2002), or Open Multilingual WordNet (Bond et al. 2014).

6. Acknowledgement

Work co-financed by EU's 7FP under grant agreement No. 316097 [ENGINE].

7. References

Bond, F., Fellbaum, Ch., Hsieh, S., Huang, Ch., Pease, A., Vossen, P. (2014). A Multilingual Lexico-Semantic Database and Ontology. In P. Buitelaar & P. Cimiano (Eds) *Towards the Multilingual Semantic Web* Paul, Springer, pp. 243–258.

Collins, A.M., Quilian, M.R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behaviour*, 8, pp. 240–247.

Deese, J. (1964). The associative structure of some common English adjectives. *Journal of Verbal Learning and Verbal Behaviour, 3* (5), pp. 347–357.

Miller, K. J. (1998). Modifiers in WordNet. In Ch. Fellbaum (Ed). *WordNet: An Electronic Lexical Database.* MIT Press: Cambridge, Massachusetts, pp. 47–68.

Maziarz, M., Szpakowicz, S., Piasecki, M. (2012). Semantic Relations among Adjectives in Polish WordNet 2.0: A New Relation Set, Discussion and Evaluation. *Cognitive Studies,* 12, pp. 149–179.

Maziarz, M., Piasecki, M., Rudnicka, E., Szpakowicz, S. (2013). Beyond the Transfer-and-Merge WordNet Construction: plWordNet and a Comparison with WordNet. In G. Angelova (Ed.) I*nternational Conference on Recent Advances in Natural Language Processing (RANLP 2013).* Association for Computational Linguistics, pp. 443–452.

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

Rudnicka, E., Maziarz, M., Piasecki, M., Szpakowicz, S. (2012). A Strategy of Mapping Polish WordNet onto Princeton WordNet. In M. Kay & Ch. Boitet (Eds.)

Proceedings of COLING 2012: Technical Papers. COLING 2012 Organizing Committee, pp. 1039–1048 .

Rudnicka, E., Witkowski, W., Kaliński M. (2015). A Semi-Automatic Adjective Mapping between plWordNet and Princeton WordNet. In P. Kral & V. Matoušek (eds.) *Text, Speech, Dialogue. 18th International Conference, TSD 2015, Pilsen, Czech Republic, September 14-17, 2015, Proceedings*. [Lecture Notes in Artificial Intelligence Series]; pp. 360–368.

Sheinman, V., Fellbaum, C., Julien, I., Schulam, P., Tokunaga, T. (2013). Large, huge or gigantic? Identifying and encoding intensity relations among adjectives in WordNet. *Language Resources and Evaluation* 47, pp. 797–816.

Vossen, P. (ed.) (2002). *EuroWordNet General Document,* Version 3 (final) URL: http://www.hum.uva.nl./~ewn