Ruled-based, Interlingual Motivated Mapping of plWordNet onto SUMO Ontology

Paweł Kędzia, Maciej Piasecki

Institute of Informatics, Wrocław University of Technology,
{pawel.kedzia, maciej.piasecki}@pwr.edu.pl

Abstract

In this paper we study a rule-based approach to mapping plWordNet onto SUMO Upper Ontology on the basis of the already existing mappings: plWordNet – Princeton WordNet – SUMO. Data acquired from the inter-lingual relations between plWordNet and Princeton WordNet and the relations between Princeton WordNet and SUMO ontology are used in the proposed rules. Several mapping rules together with the matching examples are presented. The automated mapping results were evaluated in two steps: first, (i) we automatically checked the formal correctness of mappings for the pairs of plWordNet synset and SUMO concept, next, (ii) a subset of 160 mapping examples was manually checked by two+one linguists. We analyzed types of mapping errors and their causes. The proposed rules had a very high precision, especially when the errors in the resources are taken into account. Both wordnets were constructed independently and in the effect the obtained rules are non trivial and they reveal a member of differences between the two wordnets and the two languages.

Keywords: plWordNet, SUMO, Ontology, wordnet to ontology mapping

1. Introduction

A wordnet describes lexical meanings in terms of lexico-semantic relations. Naturally, such relations are imprecise and their structures reflect the natural origin of the lexical system, e.g. linguistic hypernymy does not form a proper taxonomy system. Still, a precise semantic representation of lexical meanings, e.g. in terms of a logic-based semantic representation, is often required in language processing. A mapping of a wordnet to a formal ontology can be a step in this direction. Princeton WordNet (henceforth, PWN) (Fellbaum, 1998) – the largest wordnet in the world – has been manually mapped to Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2001; Pease, 2011) – a top level ontology. SUMO describes concepts by conceptual relations and logical expressions that can be used as a basis for the formalised semantic representation. Most existing wordnets were built following the transfer method (Vossen, 1998), so they were directly translated from PWN to a very large extent. In this way they have been automatically linked to SUMO, as their structure of synsets and synset relations mimic the structure of PWN.

However, in the case of wordnets developed independently from PWN, e.g. plWordNet[1] (Maziarz et al., 2012) – a very large wordnet of Polish, the situation is different. Differences in wordnet structures and the principles they are based on cause that an automatic transfer of wordnet-to-ontology mappings (called ontology mappings) can be difficult. For such wordnets, ontology mappings can be built manually, but this is a costly process. It seems to be harder than a manual mapping of two wordnets for different languages, because in the latter case, we can use bilingual dictionaries while in the former case the precise formal definitions of ontological concepts require a careful analysis of formal ramifications of each mapping links added. Still, the ontology mapping process can be supported by, or even based on, the existing wordnet-to-PWN mappings. The question is, how far such bilingual semantic mapping preserves formal interpretation of lexical meanings in terms of their ontological descriptions.

Our goal is to explore the possibility of utilising the existing wordnet-to-wordnet mappings in the process of mapping plWordNet to SUMO ontology. Moreover, as both PWN and plWordNet were built independently, we expect that the mapping process would not be trivial and the results may provide interesting insights on the differences between these two resources.

2. Task – mapping sequence

plWordNet is a very large wordnet for Polish which has been built since 2005. plWordNet was meant to provide a faithful description of the Polish lexical system, so neither was it translated from Princeton WordNet (transfer method (Vossen, 1998)), nor any other form of direct information transfer via bilingual resource was utilised during its entire development. plWordNet was built from scratch following the *corpus based wordnet development process* (Piasecki et al., 2013b) that was proposed as a variant of the merge method. The most important source of lexical knowledge was a huge corpus as well as tools for corpus browsing and automated extraction of lexico-semantic relations. However, the final decision for every plWordNet element is always in the hands of the lexicographers who also can consult paper dictionaries and refer to their language competence (Piasecki et al., 2013b). As a result, plWordNet and Princeton WordNet are rare examples of two large wordnets developed completely independently from each other. Moreover, the construction of plWordNet is based on the *minimal commitment principle* (Maziarz et al., 2013), according to which lexical units (pairs: lemma plus sense number) are the basic building blocks and synsets are defined as sets of such lexical units that share constitutive lexico-semantic relations (e.g. hypernymy or meronymy) and features (e.g. stylistic register, aspect or semantic verb

[1] `http://plwordnet.pwr.wroc.pl`

class). Synset relations are derived from lexico-semantic relations linking lexical units belonging to synsets. Linguistic substitution tests are an important part of relations definitions and are consulted during wordnet editing. Thus, plWordNet structure is thoroughly motivated by linguistic knowledge and language data observed in corpora.

For the purpose of work presented here, we used the latest plWordNet version 2.19 which includes ≈142 300 lemmas (literals), described with ≈200 500 lexical units (pairs: lemma, sense id) and ≈146 200 synsets. THe noun part of plWordNet has been manually mapped onto PWN 3.0 with the help of 7 interlingual semantic relations inspired by EuroWordNet relations (Vossen, 2002). The relations are organised into an ordered list and linguists were asked to choose only the first matching one: I-synonymy, I-inter-register synonymy[2], I-hyponymy, I-hypernymy, I-meronymy, I-holonymy and I-partial synonymy[3] (Rudnicka et al., 2012). The relations express different forms of dependency between the denotations of linked synsets. The sequence is ordered in such way that the relation that seems to be the most informative for the mapping applications goes before to the less informative one. The mapping process goes in plWordNet to PWN direction, so PWN synsets have no incoming inter-lingual links.

The main criterion for linking is the similarity of wordnet graphs around the source and target synsets (that are considered for linking). Relations of the linked synsets should match, because according to the plWordNet model, sets of lexico-semantic relation links are the primary tool of the lexical meaning description in a wordnet. During the comparison of wordnet structures, linguists consult translations in bilingual dictionaries and compare glosses in PWN and comments in plWordNet, whenever they are available. So far the mapping has been built for about 82 000 synsets a large subset of the noun part (≈79 600 synsets) and a small subset of the adjective part (≈2 400 synsets) of plWordNet. PWN had been earlier mapped manually onto SUMO (Niles and Pease, 2003). The mapping was transferred to several other wordnets derived from PWN. This wordnet-to-ontology mapping is based on three relations:

1. *equivalent* – expressing that a PWN synset is equivalent to a SUMO concept with respect to the synset's meaning where the main focus is given to synset denotation, e.g. *plant 2* is the equivalent of a SUMO concept *Plant*.

2. *instance of* – a PWN synset denotation is an instance of a SUMO concept, e.g. *Aristotle 1* is an instance of *Man*, or a synset denotation is a member of a collection denoted by the SUMO concept, e.g. {*Eden 2*} is a member of *Region*, or a PWN synset represents an individual entity (mostly a Proper Name) which is an instance of SUMO concept,

3. *subsumed* – is a subsumption relation between a PWN synset denotation and a SUMO concept denotation – an analogue of linguistic hyponymy – the SUMO concept is a 'hypernym' of the PWN synset, e.g. the {*town 1*} synset is linked by *subsumption* to the concept *City*.

SUMO contains definitions of general terms and facts and may be the basis for more specific ontologies (middle ontologies and/or domain ontologies).

The interlingual links from plWordNet to PWN can guide us in transferring ontological mappings to back plWordNet. When a plWordNet synset is mapped to a PWN synset by *I-synonymy* that is next mapped to a SUMO concept by the *equivalent* relation, the situation seems to be clear – the resulting mapping of a plWordNet synset to a SUMO concept should be set to the *equivalent* relation. In a large sample, we found for this simple rule no other errors than errors originating from the component mappings. However, such straightforward cases as *I-synonymy – equivalent* are in minority. *I-synonymy* corresponds to $\approx 31.85\%$ of inter-lingual relations of plWordNet 2.1[4], the *equivalent* relation accounts for $\approx 19.81\%$, they both coincide only in $\approx 3.45\%$. The other combinations of mapping links are often ambiguous with respect to possible resulting mapping, e.g. in the case of *I-hyponymy – subsumed* pattern the resulting relation is mostly *subsumed*, but not necessarily to the same SUMO concept.

3. Mapping plWordNet to SUMO Ontology

SUMO concepts are described by labels equal to English words and expressions and are commented by short textual descriptions. However, the concepts are abstract entities, the core of their definitions are formal expressions and we cannot fully trust that linguistic meanings of labels are correct descriptions of concepts. The comments are not the part of their definitions.

Wordnet mapping algorithms, e.g. (Daudé et al., 2000; Daudé et al., 2003), as well as ontology-to-ontology mapping algorithms (Hovy et al., 2013) use labels and text descriptions of synsets or concepts to find potential targets for inter-resource mappings. The results can server the basis for manual mapping, as our experience in building plWordNet mapping showed (Kędzia et al., 2013), but we can have only limited trust in automatically generated suggestions for mapping links. As already mentioned, the problem gets even bigger in the case of the application of such algorithms to formalised ontologies as the mapping targets.

Besides glosses that are provided for only a small subset of plWordNet synsets, wordnet synsets of both, plWordNet and PWN, are described by:

- coarse-grained wordnet domains that correspond to "lexicographic files" of PWN,

- hypernyms on different levels defining narrow semantic classes,

- synsets linked by different relations describing the semantic contexts.

[2] I-inter-register synonymy links two synsets whose lexical units represent two different stylistic registers (Rudnicka et al., 2012).

[3] I-partial synonymy is used for synsets whose denotations and hyponyms are overlapping.

[4] I-hyponymy covers 51.4% and I-hypernymy represents 7.8% relations.

SUMO concepts are primarily described by their superconcepts on different levels of the concept hierarchy and formal expressions provided.

On the basis of the above mentioned descriptive features, we constructed a set of 71 mapping disambiguation rules. First, we concentrated on exploring the information encoded in the pairs of the mapping links. For instance, the rule *R1* presented in Algorithm 1, expresses the aforementioned case: *I-synonymy–equivalent*. *R(PLWN_PWN)* is an interlingual relation between plWordNet and PWN, *R(PWN_SUMO)* is a relation between PWN and SUMO, and, finally, *R(PLWN_SUMO)* is the resulting mapping link set directly between a plWordNet synset and a SUMO concept. *R2* and *R3* express the observation that also *instance of* and *subsumed* links can be copied via *I-synonymy*.

Algorithm 1 Example of three simple mapping rules R1, R2 and R3.

```
R1:
if R(PLWN_PWN) = synonymy and R(PWN_SUMO) = equivalent then
    R(PLWN_SUMO) = equivalent
end if

R2:
if R(PLWN_PWN) = synonymy and R(PWN_SUMO) = instance of then
    R(PLWN_SUMO) = instance of
end if

R3:
if R(PLWN_PWN) = synonymy and R(PWN_SUMO) = subsumed then
    R(PLWN_SUMO) = subsumed
end if
```

However, link-copying rule does not always work correctly in the case of *I-synonymy–subsumed* pairs, where the SUMO concept is *EthnicGroup*. To give an example for {*Aramejczyk 1* 'Aramean'}–*subsumed–EthnicGroup* the resulting link should be *instance of*, because *Aramean* is an *instance of EthnicGroup*. Also the combination *I-synonymy–equivalent* can produce errors, e.g. in the case of the concrete numbers linked to SUMO concept *Number*, like {*dziewiątka 3* 'nine'}–*equivalent–Number*.

plWordNet synsets that are more specific than the closest PWN counterparts are linked to them by *I-hyponymy* relation. If the mapping between PWN and SUMO is *equivalent*, then we set the resulting mapping link to *subsumed*, e.g. {*tysiąc 2* 'thousand'} has been linked to *Number* by *subsumed*.

Yet, such simple approach based on interlingual relations and relations between Princeton WordNet and SUMO is not sufficient. In many cases we referred to wordnet domains of synsets (of plWordNet and/or PWN). In both rules *R4** and *R5** presented in Algorithm 2, we used the domain of a plWordNet synset. In *R4** and *R5** besides analysing the interlingual relations and mappings between PWN and SUMO, we also check the domain of a plWordNet synset (*D(PLWN)*). In *R4** we test if the domain is one of: {czc, grp, jedz, msc, por, pos, rsl, rz, sbst, umy, wytw, zdarz, zj, zwz} [5] and in *R*5* whether the domain is included in the set: {grp, msc, umy, os}.

Algorithm 2 Example of mapping rules uses domain of synsets from plWordNet

```
R4*:
if R(PLWN_PWN) = partial synonymy and R(PWN_SUMO) = equivalent and D(PLWN) $\in \{czc, grp, jedz, msc, por, pos, rsl, rz, sbst, umy, wytw, zdarz, zj, zwz\}$ then
    R(PLWN_SUMO) = subsumed
end if
if R(PLWN_PWN) = partial synonymy and R(PWN_SUMO) = equivalent and D(PLWN) $\in \{os\}$ then
    if PLWN_SYNSET starts with upper letter then
        R(PLWN_SUMO) = subsumed
    else
        R(PLWN_SUMO) = manually
    end if
end if

R5*:
if R(PLWN_PWN) = partial synonymy and R(PWN_SUMO) = instance of then
    if D(PLWN) $\in \{grp, msc\}$ then
        R(PLWN_SUMO) = subsumed
    end if
    if D(PLWN) $\in \{os, umy\}$ then
        R(PLWN_SUMO) = instance of
    end if
end if
```

More complicated rules, e.g. *R6** presented in the Algorithm 3, use information about plWordNet and PWN synset domains. When plWordNet and PWN synsets are linked by *I-part of-meronymy* relation and PWN is mapped onto SUMO with *equivalent* relation, and when a plWordNet synset starts with a capital letter, then:

1. When the domain of plWordNet is *msc* and the domain of PWN is *rz*, then the relation between plWordNet and SUMO is *instance of.*

2. When the domain of plWordNet is *zwz* and the domain of PWN is *msc*, then the relation between plWordNet and SUMO is *instance of.*

In the first case, the example is plWordNet synset {*Azja Środkowa 1 (msc)* 'Central Asia'} connected on *I-part of-meronymy* to PWN synset {*Asia 1 (r)*}, and PWN synset is connected by *equivalent* to the SUMO *Asia*. So, the *R6** identifies the mapping relation as the *instance of.* For the second case, the example can be {*Hetmańszczyzna 1 (zwz)* '≈all things associated

[5] czc – body part, grp – group, jedz – food, msc – location, por – communication pos – possesing, rsl – plants, rz – things, sbst – substances, umy – connected to thinking, wytw – artifacts, zdarz – events, zj – natural phenomena, and zwz – associations

Algorithm 3 Example of mapping rules uses domains of synsets from plWordNet and PWN

```
R6*:
if R(PLWN_PWN) = partial meronymy and R(PWN_SUMO) = equivalent then
    if PLWN_SYNSET starts with upper letter then
        if D(PLWN) $\in \{msc\}$ and D(PWN) $\in \{rz\}$ then
            R(PLWN_SUMO) = instance of
        end if
        if D(PLWN) $\in \{zwz\}$ and D(PWN) $\in \{msc\}$ then
            R(PLWN_SUMO) = instance of
        end if
    end if
end if
```

with hetman'}–*I-part of-meronymy*–{*Ukraine 1 (msc)*}–*equivalent*–*Ukraine*. Finally, based on the *R6** mapping relation between {*Hetmańszczyzna 1 (zwz)*} and *Ukraine* is *instance of*.

Among the proposed 80 rules there are rules that use information about the type of SUMO concept arising from PWN-SUMO mapping. In the Algorithm 4 there are presented rules of the type *R7** in which the type of SUMO concept is checked. When the *I-partial-synonymy* and PWN-SUMO mapping function is *subsumed*, and when the domain of plWordNet synset is *os*, then:

1. If plWordNet synset does not start with a capital letter, then plWordNet-SUMO mapping relation is *subsumed*, for example: {*relacja 1 (por)* 'relation'}–*I-partial-synonymy*–{*report 1 (por)* 'report'}–*subsumed*–*Report* → plWordNet-SUMO relation is *subsumed*.

2. If a SUMO concept is one of $\{EthnicGroup, Agent\}$, then plWordNet-SUMO mapping relation is *instance of*, for example: {*Pigmej 1 (os)*}–*I-partial-synonymy*–{*Pygmy 2 (os)*}–*subsumed*–*EthnicGroup* → plWordNet-SUMO relation is *instance of*. Another example is {*Osoba Boska 1 (os)*}–*I-partial-synonymy*–{*hypostasis 3 (os)*}–*subsumed*–*Agent* → plWordNet-SUMO relation is *instance of*.

4. Results and evaluation

For the experiments we used plWordNet 2.19 including more than 82 500 interlingual links to PWN and the links defined between PWN and SUMO. As a result, we extracted 82 008 triples {plWordNet synset, Princeton WordNet synset, SUMO concept} , e.g. ({*statek 1* 'ship'}, {*boat 1*}, *WaterVehicle*). In the second step we processed the triples using all available additional information (wordnet domains and mapping links – interlingal mappings, and mappings between PWN and SUMO) in order to identify the resulting mapping relation between the processed plWordNet synset and the SUMO concept, e.g. {*statek 1*}–*subsumed*–*WaterVehicle*).

Algorithm 4 Example of mapping rule uses information about the SUMO concept

```
R7*:
if R(PLWN_PWN) = partial synonymy and R(PWN_SUMO) = subsumed then
    if D(PLWN) $\in \{os\}$ then
        if !PLWN_SYNSET starts with upper letter then
            R(PLWN_SUMO) = subsumed
        end if

        if SUMO_CONCEPT $\in \{EthnicGroup, Agent\}$ then
            R(PLWN_SUMO) = instance of
        end if
    end if
end if
```

After the second step we were left with 4 397 triples for which we could not decide about the resulting mapping relation, e.g. in the case of the triple: ({*Berno 1*}–*I-meronymy*–{*Switzerland 1*}–*equivalent*–*Switzerland*) {*Berno 1*} is not an instance of *Switzerland* and a relation closer to meronymy was not considered by (Niles and Pease, 2003). Because a synset may have several relations with the SUMO concepts, we analysed the situations where the synset is not mapped to any concept. In Table 1 we show the numbers of the mapped and not mapped unique plWordNet synsets, processed by our system.

Table 1: Number of synsets mapped to plWordNet and the remaining ones (not mapped yet).

POS	**NM**	**Mapped**	**NM [%]**
Noun	2265	70045	3
Verb	1	13	7
Adjective	141	2508	5

The last column contains the percentage of synsets that were not mapped in relation to the concrete Part of Speech. For example, our algorithm did not assign a relation for the following noun synsets:

- {arena 4 (msc) 'arena'}–*I-part of-meronymy*–{circus 5 (wytw)}–*subsumed*–Region
- {krwiobieg płucny 1 (czc) 'pulmonary bloodstream'}–*I-part of-meronymy*–{circulatory system 1 (czc)}–*subsumed*–BodyPart
- {bombeczka 1 (wyt) 'small bomb'}–*I-part of-meronymy*–{Christmas tree 5 (wytw)}–subsumed–Plant
- {kilokaloria 2 (il) 'kilo-calorie'}–*I-part of-holonymy*–{calorie 2 (il)}–*equivalent*–Calorie

The plWordNet synset {*arena 4 (msc)*} is a meronym of the PWN synset {*circus 5 (wytw)*}. The PWN synset is connected by *subsumed* relation to the SUMO concept *Region*. The algorithm did not assign any relation, but the *subsumed* link should be added for {*arena 4 (msc)*} and *Region*, because {*arena 4 (msc)*} is the place in the {*circus 5 (wytw)*}. A similar situation is observed in the case of {*krwiobieg płucny 1 (czc)*}. This synset should be connected to *BodyPart* concept with *subsumed* relation, because {*krwiobieg płucny 1 (czc)*} is the part of the {*circulatory system 1 (czc)*}. However, the decision for {*bombeczka 1 (wyt)*} was not good, because there is no a relation that could occur between {*bombeczka 1 (wyt)*} and the concept *Plant*. The same situation is in the case of {*kilokaloria 2 (il)*}: between {*kilokaloria 2 (il)*} and {*calorie.2 (il)*} holds the *holonymy* relation which means that the {*calorie 2 (il)*} is a part of {*kilokaloria 2 (il)*}, and between {*calorie 2 (il)*} and SUMO concept *Calorie* holds *equivalent* relation. The basic set of mapping relations does not cover holonymy/meronymy relations, but such a rule should appear in this case.

In the first step of evaluation we automatically checked if there are plWordNet synsets for which two or more different resulting links to the same SUMO concept were created. We found 89 such cases involving 44 plWordNet synsets, i.e. about 0.06% of all mapping decisions. For instance plWordNet synset {*igła 1* 'needle'} is linked by *I-synonymy* to PWN {*leaf 1*} and by *I-synonymy* to PWN {*acerate leaf 1*}. On the basis the mapping relations between PWN synsets and SUMO concepts: {*leaf 1*}*–equivalent–Leaf* and {*acerate leaf 5*}*–subsumed–Leaf*) our algorithm mapped {*igła 1*} onto *Leaf* using two different relations: *equivalent* and *subsumed*. A similar situation can be noticed for the synset {*rodzina 3*}, which is connected by *I-synonymy* with the PWN synsets {*kin 2*} (mapped to SUMO with *equivalent* relation) and {*family 4*} (mapped to SUMO with *subsumed* relation). Both PWN synsets have been mapped onto the same *FamilyGroup* concept, so there are two resulting plWordNet synset mappings to *FamilyGroup*: *subsumed* and *equivalent* relations

For manual evaluation we randomly selected a sample of 160 mapping decisions. Each selected mapping decision was verified by two linguists and classified as correct, linked to a hypernym of the appropriate concept and incorrect. The same classes were used for all types of ontological links. As a result, ≈83,13% links were verified as correct, a few were partially correct, and 15% were plainly incorrect. In Table 2 there is a sample of mappings followed by the information about the evaluator agreement and the final decisions. The column *L1* contains the decision of the first linguist, *L2* contains the decision of the second linguist and the *Decision* column includes the final decision. When the decisions of *L1* and *L2* differ, the final decision was made by the third linguist. In the examples, *eq* means that between plWordNet synset and SUMO concept holds *equivalent* relation, *io – instance of*, and *sbs subsumed*. The agreement between linguists *L1* and *L2*, was 81%.

Table 2: Examples of linguist decisions.

Example	L1	L2	Decision
sobota.1 -eq -Saturday	t	t	t
Włoch.1 -io- EthnicGroup	t	t	t
brzdęk.1 -sbs- RadiatingSound	t	t	t
obiór.1 -eq- Election	t	f	f
Cyganeczka.1 -io - EthnicGroup	f	t	t
szurpek.1 -sbs- Moss	t	f	t
prawnik.1 -eq- FieldOfLaw	f	f	f
mason.1 -io- Organization	f	f	f
naprawiciel.1 -sbs- Position	f	f	f

Table 3 shows the number of plWordnet synsets in relation to Part of Speech and the mapping relation type to SUMO concept. The column *eq.* represents the *equivalent* relation, *i.o. – instance of*, *sbs. subsumed* and the column *nm* contains number of *not mapped* synsets. Not mapped synsets are synsets for which our algorithm could not assign a relation to the pair $< plWordnetSynset, SUMOConcept >$ to be mapped.

Table 3: The number of synets with respect to the part of speech and relation type between plWordnet synset and SUMO concept.

POS	Relation			
	eq.	i.o.	sbs.	nm
Noun	2295	2464	70416	4256
Adjective	133	194	2093	141
Verb	2	1	1	1

Table 4 contains information about five most frequent domains of plWordnet synsets in the mapping process in relation to the resulting mapping relations. The columns *cnt* contains the number of involved domains and the columns *dom* contains information about the domains. It can be noticed that the synsets with *wytw* domain were the most commonly mapped with the *equivalent* and *subsumed* relations. In the case of the *instance of* relation, the most commonly mapped domain of synset was *umy*.

Table 4: The most frequent domains in the result of mapping for each mapping relation.

Relation							
eq.		i.o.		sbs.		nm.	
dom.	cnt	dom.	cnt.	dom.	cnt.	dom.	cnt.
wytw	344	umy	849	wytw	11775	os	1090
msc	246	os	685	os	11347	wytw	604
pos	182	il	274	zw	8549	msc	529
sbst	175	msc	252	rsl	5276	sys	387
por	159	rel	176	msc	3765	grp	372

Table 5 includes statistics about the five most frequent SUMO concepts used in the mapping plWordnet to SUMO in the context of each relation. On the basis of those statistics we can notice that *SubjectiveAssessmentAttribute* is the most frequently used concept in the case of *equivalent* relation. These mappings are incorrect, but a lot of them are the result of the incorrect mapping PWN to SUMO. For instance, {*porażka 1 (zdarz)* 'failure'}*–I-synonymy–*{*failure 2 (zdarz)*}*–equivalent–SubjectiveAssessmentAttribute* – based on the *interlingual synonymy* between {*porażka 1 (zdarz)*} and {*failure 2 (zdarz)*}, the rule copy only the PWN-SUMO relation, which is *equivalent* to *SubjectiveAssessmentAt-*

tribute concept. The same situation happens in the case of: {*bohater 5i (os)* 'hero'}*–I-synonymy–*{*hero 1 (os)*}*–equivalent–SubjectiveAssessmentAttribute*. The most frequent concept used with *instance of* relation is *FieldOfStudy*. There is one interesting example where the same plWordNet synset is mapped by different I-relations to three different PWN synsets:

Table 5: The number of five most frequent SUMO concepts in the mapping results for each relation.

Relation	Concept	Count
equivalent	SubjectiveAssessmentAttribute	58
	Currency	11
	LengthMeasure	10
	GroupOfPeople	10
	Woman	8
instance of	FieldOfStudy	777
	EthnicGroup	455
	UnitofMeasure	140
	Man	91
	Nation	90
subsumed	FloweringPlant	4123
	SubjectiveAssessmentAttribute	3015
	Human	2873
	Position	2461
	Device	2251

- {*algebra 1 (umy)* 'algebra'}*–I-part of-holonymy–*{*quadratics 1 (umy)*}*–instance of–FieldOfStudy*
- {*algebra 1 (umy)*}*–I-part of-holonymy–*{*vector algebra 1 (umy)*}*–instance of–FieldOfStudy*
- {*algebra 1 (umy)*}*–I-synonymy–*{*algebra 1 (umy)*}*–instance of–FieldOfStudy*

and as the result {*algebra 1 (umy)*} is correctly mapped to *FieldOfStudy* with *instance of* relation.

In the third group, there are mappings in which the relation between plWordNet synset and SUMO concept is *subsumed*. *FloweringPlant* is the most frequent target concept in this group. Below we present examples of correct mappings with *subsumed* relation:

- {*konopie 3 (rsl)* 'cannabis'}*–I-synonymy–*{*cannabis 1 (rsl)*}*–subsumed–FloweringPlant*; on the basis of these information, the final relation between {*konopie 3 (rsl)*} and *FloweringPlant* is *subsumed*.
- {*kozibród 1 (rsl)*}*–I-meronymy–*{*Tragopogon 1 (rsl)*}*–subsumed–FloweringPlant*; the relation between {*kozibród 1 (rsl)*} and *FloweringPlant* is *subsumed*.

5. Error analysis

After resolving the differences between the two annotators we went through all the errors confirmed and we analysed their potential causes. Most errors were caused by errors that had been made in both mappings we used, namely, plWordNet-to-PWN (inter-lingual) and PWN-to-SUMO (ontological) mappings. Concerning the inter-lingual mappings, the mapping errors resulted from the wrong sense considered on one of sides or, more often, too general English synset selected in the case of the I-hyponomy relation. For example, the plWordNet synset {*półmetek* 2, *połowinki* 1} – both means a kind of party organised in the middle of school or study period – was mapped by I-hyponymy to the English synset {*social event* 1}, while it should be linked by I-hyponymy to {*party* 4} which is an indirect hyponym of the {*social event* 1}. This erroneous, too general mapping caused that our rules suggested *SocialInteraction* as the SUMO concept subsuming the Polish synset. It is too general, *SocialParty* is a proper choice, and it would have been suggested by a correct inter-lingual mapping.

Errors in the ontological mapping seem to be mostly caused by applying some schematic way of adding mapping links to too many words that are too specific and do fit to this scheme. For instance, the Polish synset {*naprawiciel* 1 'humorously, someone who have repaired something'} is linked by *inter-register synonymy* to another Polish synset {*reperator* 1, *naprawca* 1} 'the same meaning but without this sense of humor' which has been mapped to the English {*mender* 1, *repairer* 1, *fixer* 3} by I-synonymy relation. The English synset was linked by *subsumption* to *Position*. However, the formal definition of the *Position* concept requires that an instance of this concept must have a specified role in some organisation. Workers denoted by the synset {*mender* 1, ...} do not have such a role assigned. As a result, the mapping of the Polish synset to *Position* sounds funny.

The Polish synset {*schola* 1} '≈small amateur church choir' was linked by I-hyponymy to {*chorus* 2} which was mapped to the SUMO concept *Group*. However, we can notice that there is *MusicalGroup* down in the subhierarchy of *Group* and it is a better concept for {*chorus* 2}, as it is more specific.

In some rare cases, we noticed errors that coincide in both mappings, e.g. {*żona* 1} was mapped by I-synonymy to {*wife* 1}, but also by I-hypernymy to {*battle-ax* 2, *battle-axe* 1} ("a sharp-tongued domineering wife") that was mapped to *SubjectiveAssessmentAttribute*. We did not expect to find multiple inter-lingual mappings (the I-hypernym was in fact added as a relation in the opposite direction), but also the mapping to SUMO concept is not correct.

In a few cases, SUMO concept definition expression include some specific restrictions that does not match plWordNet synset definition in the form of the relation structure, but the concept seems to match the synset only when the synset lemmas and the concept label are compared. For instance, {*dzwon* 1 'bell'} is linked by I-synonymy to {*bell* 1} mapped by *subsumption* to *Bell*, so it looks well, but the formal definition of *Bell* constraints its denotation to musical instruments, while {*dzwon* 1} 'bell' is not a musical instrument in Polish.

Some errors were caused by a different treatment of regular polysemy in both wordnets: some senses are represented in one wordnet by different lexical units of the same lemma while in the second wordnet there is one lexical unit of a more vague sense corresponding to the merged senses of the lexical units from the first wordnet. The SUMO mapping is based on one of the two merged senses.

SUMO hierarchy in many places is very shallow. The suggested links are proper with respect to the SUMO structure, but do not provide enough information. In order to improve this, SUMO should be supplemented by a medium level ontology. For instance, {*mieczyk* 1, *glaudiolus* 1, *gladiola* 1} is linked by I-synonymy to {*genus Gladiolus* 1}, which is linked by *subsumption* to *FloweringPlant*. This ontological link is correct, but there are many different kinds of flowering plants. In the case of some SUMO concepts the sex of subclasses is small and surprisingly selective.

Only a very small percentage of errors results from flaws in the ontology mapping rules. Thus, their precision seems to be very high. As a result, the ontology mapping rules can be used to support manual verification of both plWordNet-to-PWN and PWN-to-SUMO mappings. However, it is still necessary to manually analyse automatically suggested links to SUMO concepts and then to investigate the sources of the observed discrepancies.

6. Conclusions and future works

In order to resolve mappings in some of the not yet processed triples, we plan to use hypo/hyperonymy structure in plWordNet and PWN in order to transfer the mapping to SUMO via indirect connections. For instance, plWordNet synset {*Anglik 2* 'Englishman'} is linked by *I-meronymy:element* to PWN {*English 2*}, which is mapped by *equivalent* relation to the *UnitedKingdom* SUMO concept. In this case our rules cannot take any decision about final-mapping relation, due to the I-meronymy relation. However, we can refer to the hypernyms linked to PWN by I-synonymy, in Fig. 1 {*Brytyjczyk 1* 'Britisher'} is connected by *I-synonymy* to {*Britisher 1*} and check relation between this synset to SUMO ontology. {*Britisher 1*} is mapped by *instance of* onto *EthnicGroup*. On this basis, we can add *instance of* relation from {*Anglik 2*} to *EthnicGroup* concept because {*Anglik 2*} is a hyponym of {*Brytyjczyk 1*}.

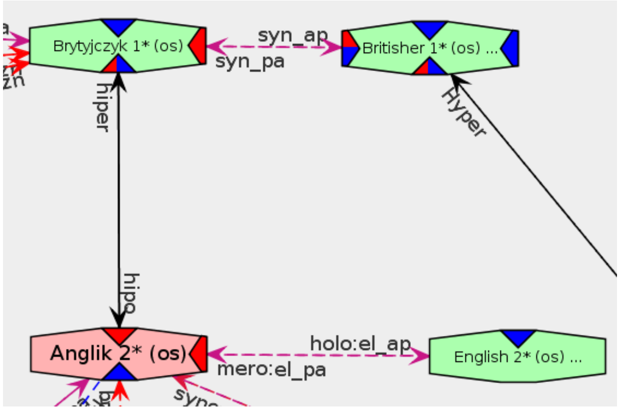

Figure 1: Hypernymy and I-synonymy relation for {*Anglik 2*} and {*Brytyjczyk 1*} synsets. A part of the screen shots from the *WordnetLoom* application (Piasecki et al., 2013a)

In the case of example from Sec. 4.: {*Berno 1*} – *Switzerland*, it is possible to map {*Berno 1*} to *Switzerland* concept by introducing a new ontological mapping relation: *meronymy*. By expanding SUMO with middle level ontologies, we could also connect larger number of the plWordNet specific synsets with the appropriate relations.

In the case of wordnets translated from PWN, it is often assumed that the mapping PWN-onto-SUMO can be automatically transferred on the wordnet linked to PWN. Our work showed that in the case of two independently constructed wordnets, this transfer is not as simple as one could expect. The task is harder due to a variety of relations linking plWordNet and PWN, but these relations are necessary for the proper descriptions of the interlingual links.

On the other hand, the results are very good, as using simple rules we were able to obtain results that are a very good basis for the construction of the complete ontology mapping. There is also room for improving the rules by exploring the structures of both wordnets and referring to narrower semantic domains.

Some encountered errors of our mapping rules originated from the errors in the wordnet-to-wordnet and wordnet-to-ontology mappings. So the rules can be also used as a diagnostic tool. The wordnet-to-ontology mapping could be also used during building the inter-lingual mapping, but formal definitions could be too problematic for linguists working on the mapping.

Acknowledgments

The work is co-financed by the Polish National Centre for Research and Development project *SyNaT* and the European Union within European Innovative Economy Programme project *NEKST* POIG.01.01.02-14-013/09.

7. References

J. Daudé, L. Padró, G. Rigau, Departament Llenguatges, and Sistemes Informhtics. 2000. Mapping wordnets using structural information. In *In Proceedings 38 th Annual Meeting of the Association for Computational Linguistics(ACL00). Hong Kong*.

Jordi Daudé, Lluiś Padró, and German Rigau. 2003. Making wordnet mappings robust. In *Proceedings of the 19th Congreso de la Sociedad Española para el Procesamiento del Lenguage Natural, SEPLN*, Universidad Universidad de Alcalá de Henares. Madrid, Spain.

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. The MIT Press.

E. Hovy, R. Navigli, and S. P. Ponzetto. 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Paweł Kędzia, Maciej Piasecki, Ewa Rudnicka, and Konrad Przybycień. 2013. Automatic prompt system in the process of mapping plWordNet on Princeton WordNet. *Cognitive Studies*.

Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. Approaching plWordNet 2.0. In Christiane Fellbaum and Piek Vossen, editors, *Proceedings of 6th International Global Wordnet Conference*, pages 189–196, Matsue, Japan, January. The Global WordNet Association. Book: http://www.globalwordnet.org/gwa/proceedings/gwc2012.pdf.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2013. The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. *Langauge Resources and Evaluation*, 47(3):769–796.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, pages 2–9, New York, NY, USA. ACM.

Ian Niles and Adam Pease. 2003. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03), Las Vegas*, pages 412–416.

Adam Pease. 2011. *Ontology: A Practical Guide*.

Maciej Piasecki, Michał Marcińczuk, Radosław Ramocki, and Marek Maziarz. 2013a. WordnetLoom: a wordnet development system integrating form-based and graph-based perspectives. *International Journal of Data Mining, Modelling and Management*, 5(3):210–232.

Maciej Piasecki, Radoslaw Ramocki, and Michal Kalinski. 2013b. Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, Hissar, Bulgaria, September 2013*, pages 553–561, Shoumen. ACL, INCOMA Ltd. and ACL.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. A strategy of mapping Polish Wordnet onto Princeton Wordnet. In *Proceedings of COLING 2012: Posters*, pages 1039–1048, Mumbai, India, December. The COLING 2012 Organizing Committee.

Piek Vossen. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publisher.

Piek Vossen. 2002. EuroWordNet general document version 3. Technical report, University of Amsterdam.