A General Methodology for Mapping EuroWordNets to the Suggested Upper Merged Ontology

Dennis Spohr

Universität Stuttgart
Institut für Linguistik/Romanistik
D-70174 Stuttgart
dennis.spohr@ling.uni-stuttgart.de

Abstract

This paper presents a general methodology to mapping EuroWordNets (Vossen, 1998) to the Suggested Upper Merged Ontology (SUMO; (Niles and Pease, 2001)), and we show its application to the French EuroWordNet. The process makes use of existing work on mapping Princeton WordNet (Fellbaum, 1998) to SUMO (Niles and Pease, 2003). After a general discussion of the usefulness of our approach, we provide details on the procedure of mapping individual EuroWordNet synsets to SUMO conceptual classes, and discuss issues arising from a fully automatic mapping. In addition to this, we present a quantitative analysis of the thus created semantic resource and discuss how the accuracy in determining the correct SUMO class for a particular EuroWordNet synset might be improved. Finally, we briefly hint at how such resources may be used, e.g. in order to extract selectional preferences of verbal predicates with respect to the ontological categories of their syntactic arguments.

1. Introduction

Semantic lexicons have become increasingly important over the last decades, with Princeton WordNet (Fellbaum, 1998), EuroWordNet (Vossen, 1998) and FrameNet (Baker et al., 1998) being probably the best-known and most widely used ones in the field of lexical-semantic natural language processing. In recent years, however, there has also been a strong tendency towards interfacing such lexical resources with *knowledge bases* or *taxonomies* of general knowledge (e.g. OntoWordNet project; Gangemi et al. (2003)), both commonly though often inaccurately referred to as *ontologies*. Well-known examples of such efforts are e.g. Vossen et al. (1998), who linked EuroWordNet's Inter-Lingual-Index to a number of *base concepts* and a *top ontology* as integral part of the EuroWordNet project, Niles and Pease (2003) who mapped Princeton WordNet to the Suggested Upper Merged Ontology (SUMO), and Scheffczyk et al. (2006) and Reiter (2007) who linked FrameNet and SUMO. Moreover, the recent Global WordNet Grid[1] is pursuing such efforts on a considerable scale to create mappings from SUMO to all existing WordNets (see also Horák et al. (2008)).

One of the main reasons for the importance of such approaches is that while resources like (Euro-)WordNet and FrameNet attempt to model lexical-semantic knowledge, ontologies try to mediate common knowledge or knowledge of the world. Therefore, linking these two types of resources may be able to bridge the gap between language-dependent lexical knowledge and language-independent ontological or world knowledge.

In this paper, we present a general methodology for mapping EuroWordNets to SUMO by using both an existing mapping from Princeton WordNet 1.6 to SUMO (Niles and Pease, 2003) and the linking of the EuroWordNets to the Inter-Lingual-Index (Vossen et al., 1998). We apply our methodology to the French EuroWordNet. Section 2. of this paper introduces some background on WordNet and SUMO, and in Section 3. we will present our methodology for mapping EuroWordNets to SUMO. After an evaluation of the mapping methodology, we conclude in Section 5. and briefly discuss ways to apply and further extend our approach.

[1] http://www.globalwordnet.org/gwa/gwa_grid.htm

2. Background

In this section, we will provide some background on (Euro)WordNet, the Suggested Upper Merged Ontology, as well as approaches to mapping the two resources.

2.1. (Euro)WordNet

The WordNet project, which was initiated at Princeton University during the 1980s (cf. Fellbaum (1998)), is certainly one of the projects that have had huge impact on the NLP community. WordNet is a lexical semantic dictionary of English whose structure is guided by psycholinguistic principles. In WordNet, lexical items are organised in so-called *synsets* – sets of semantically synonymous words – which in turn are linked by a set of lexical semantic relations, such as *hypernymy/hyponymy*, *holonymy/meronymy* and *troponymy*. The first version of WordNet was released in June 1991 and contained roughly 40,000 synsets. In its current version 3.0, WordNet contains almost 120,000 synsets for English.

In the late 1990s, the EuroWordNet project aimed at the creation of WordNets for several European languages. While the EuroWordNets retain the general organisation of WordNet using synsets as primary entities, they further linked each synset of each EuroWordNet to the so-called *Inter-Lingual-Index* (ILI; cf. Vossen (1998)) that serves as an interlingua between all EuroWordNets. Moreover, the entities in the ILI are linked to the *EuroWordNet Top Ontology*, a set of concepts inspired by Pustejovsky's *qualia* (Vossen et al. (1998); cf. Pustejovsky (1995)). After the completion of the project in 1999, the created EuroWordNets consisted of between 9,300 (for Estonian) and 48,500

```
00001740 03 n 02 entity 0 something 0 014 ~
00002086 n 0000 ~ 00003095 n 0000 ~
00003731 n 0000 ~ 00009457 n 0000 ~
03435902 n 0000 ~ 03495843 n 0000 ~
03614902 n 0000 ~ 06331805 n 0000 ~
06683928 n 0000 ~ 06684175 n 0000 ~
06846327 n 0000 ~ 06847052 n 0000 ~
06847350 n 0000 ~ 06847525 n 0000 |
anything having existence (living or
nonliving) &%Physical=
```

Figure 1: WordNet 1.6 entry of synset `00001740-n` containing the SUMO mapping

(for Italian) synsets, which are available from the European Language Resources Association at different rates[2].

2.2. Suggested Upper Merged Ontology

The Suggested Upper Merged Ontology (SUMO; Niles and Pease (2001)) is a freely available upper-level ontology owned by the IEEE. According to its official website[3], SUMO *"and its domain ontologies form the largest formal public ontology in existence today"*. SUMO is connected to the domain ontologies via the *Mid-Level Ontology* (MILO). SUMO itself is expressed in SUO-KIF, a language derived from the Knowledge Interchange Format (KIF; Genesereth (1991)), that equals first-order logic in expressivity (cf. Reiter (2007): p. 26).
The size as well as the high degree of formalisation in SUMO make it highly attractive for natural language processing, and thus there have been approaches to mapping SUMO to NLP lexicons, such as FrameNet (Baker et al., 1998) and WordNet. In the following section, we will focus on the latter of these approaches.

2.3. Mapping WordNet and SUMO

Niles and Pease (2003) have created a manual mapping from version 1.6 of WordNet to SUMO, and have in subsequent years released new mappings for each new version of WordNet. In creating their linking, Niles and Pease (2003) have decided to use the following three mapping relations: *synonymy* (equivalence '=') indicates that the WordNet synset is equivalent to the SUMO concept, *hypernymy* (subclass-superclass relation '+') indicates that the synset is a hyponym of (i.e. more specific than) the SUMO concept, and *instantiation* ('@') means that the synset is an instance of the respective SUMO concept (for more details see Niles and Pease (2003): p. 413). The mapping information is simply added to the existing WordNet database, which yields a structure like the one for the synset containing *entity* and *something* shown in Figure 1, with `&%Physical=` at the end of the line indicating the mapped SUMO concept (`Physical`) as well as the mapping relation ('='). For the first mapping created by Niles and Pease (2003) for version 1.6 of WordNet, 59,550 noun mappings contained the '+', 5,575 the '@' and 947 the '=' relation, while 12,019 of the verb mappings contained the '+' and 108 the '=' relation.

[2]See http://catalog.elra.info/ for details.
[3]http://www.ontologyportal.org/

As was mentioned in the introduction, the Global WordNet Grid initiative, which was launched in early 2006, is trying to provide WordNet-SUMO mappings for all existing WordNets. The current state, as of early 2008, comprises mappings for 5,000 English base concepts, as well as for the Spanish and Catalan WordNets.
In contrast to the approaches presented so far, we try to create mappings fully automatically, though building on the previous manual work of Niles and Pease (2003).

3. Mapping EuroWordNet to SUMO

In this section, we will present how the French EuroWordNet has been mapped onto SUMO conceptual classes. The general methodology of creating the mapping to the French EuroWordNet is described in the following subsection.
Although there exists a recent mapping to version 3.0 of WordNet, we decided to use the first mapping of WordNet 1.6 as starting point. The reason for doing so is discussed below, which deals with the sensemaps between different versions of WordNet (see Section 3.2.).

3.1. General methodology

As is the case with all EuroWordNets, the French EuroWordNet is linked to the Inter-Lingual-Index, a set of concepts that is intended to be largely language-independent (cf. Vossen et al. (1998)). A crucial prerequisite for our approach to function is that the identifiers of entities in the Inter-Lingual-Index correspond to synset identifiers in version 1.5 of WordNet. For example, entity `00058624-n` of the Inter-Lingual-Index, which is glossed by *"the launching of a rocket under its own power"*, corresponds to synset {`décollage_1`, `lancement_d'une_fusée_1`} in EuroWordNet French and to {`blastoff_1`, `rocket_firing_1`, `rocket_launching_1`, `shoot_1`} in WordNet 1.5. Starting from these observations, i.e. the mapping of SUMO to WordNet 1.6 and the linking of the French EuroWordNet to the Inter-Lingual-Index ($\approx$ WordNet 1.5), the remaining task is to move from WordNet 1.5 to 1.6. For this, we can avail ourselves of the sensemap files that came with the 1.6 release of WordNet, which indicate the changes from version 1.5 to 1.6. Ignoring particular issues for the moment (see Section 3.2.), the resulting EuroWordNet entries look like the one shown in Figure 2. The structure is based on the format suggested by the Global WordNet Grid. The mapping process is summarised in Figure 3.

3.2. WordNet sensemaps

Whenever updates of WordNet are released, the updated version comes with files that, among others, indicate changes in the structure of the synsets. For example, synset `00058624-n` from above has been split in the step from WordNet 1.5 to 1.6: {`shoot_1`} is now a member of synset `00078261-n`, {`blastoff_1`} of synset `00065319-n`, and {`rocket_firing_1`, `rocket_launching_1`} of synset `00065148-n`. Therefore, version 1.6 contains new synsets that did not exist in version 1.5, and further cases in which a synset is reorganised thus that some of its items belong to different synsets in the updated version. The primary problem for the task of

```
<SYNSET>
  <POS>n</POS>
  <SYNONYM>
    <LITERAL>organisme
      <SENSE>1</SENSE>
    </LITERAL>
    <LITERAL>forme de vie
      <SENSE>1</SENSE>
    </LITERAL>
    <LITERAL>être
      <SENSE>2</SENSE>
    </LITERAL>
    <LITERAL>vie
      <SENSE>11</SENSE>
    </LITERAL>
  </SYNONYM>
  <ILI>00002728-n</ILI>
  <HYPERONYM>00002403-n</HYPERONYM>
  <SUMO>Organism
    <TYPE>=</TYPE>
  </SUMO>
  <DEF>any living entity</DEF>
</SYNSET>
```

Figure 2: EuroWordNet entry of synset `00002728-n` after the mapping

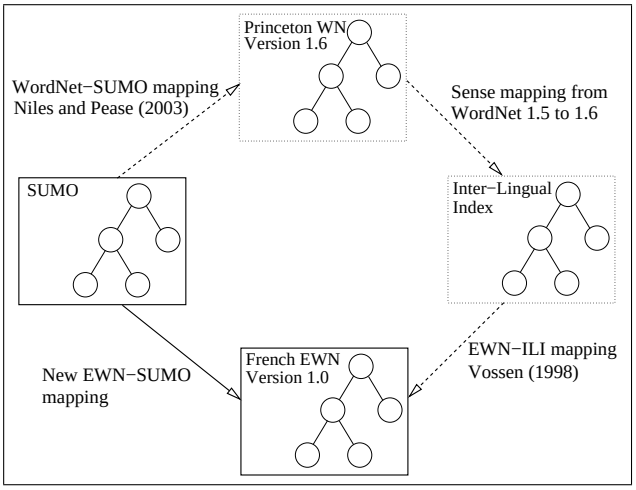

Figure 3: Process of mapping the French EuroWordNet to SUMO (clockwise from left-hand side)

mapping such instances comes from the fact that individual members of a synset do not have unique identifiers themselves, but only the synset as a whole[4]. Therefore, when a synset has been split, it is not possible to automatically determine the correct position at which the synset has to be split in a different language, or even whether it has to be split at all. Moreover, each update comes with a large number of such changes, and therefore using the most recent mapping between SUMO and WordNet 3.0, which is without a doubt desirable, would multiply the inaccuracies

[4]This is, of course, not a problem of the WordNet approach, but rather of the fact that there is no one-to-one mapping between languages.

```
<SYNSET>
    ...
  <ILI>00058624-n</ILI>
  <HYPERONYM>00058381-n</HYPERONYM>
  <SUMO>Impelling
    <TYPE>+</TYPE>
  </SUMO>
  <SUMO>Motion
    <TYPE>+</TYPE>
  </SUMO>
  <SUMO>Shooting
    <TYPE>=</TYPE>
  </SUMO>
</SYNSET>
```

Figure 4: Part of the EuroWordNet entry of synset `00058624-n` ({`décollage_1`, `lancement_d'une_fusée_1`}) after the mapping

in the mapping right from the start. Just imagine a case in which a synset has been split e.g. from WordNet 1.5 to WordNet 1.6, and the new synset is then split again when going to 1.7, and so on[5].

The decision that was made for cases like these is to assign to the original synset two (or more if necessary) SUMO classes: first the one that has been mapped to this synset, and second the ones to which the new (or relevant existing) synsets have been mapped in WordNet 1.6. The justification of this decision is based on the assumption that on a level as abstract as that of SUMO conceptual classes, a "slight" reorganisation of the synsets and some of their items should not lead to significant conceptual clashes, as this would imply that grave errors had been made when putting the respective senses into one synset in the first place. In Figure 4 above, which depicts the entry of synset `00058624-n` after the mapping, we see that the two SUMO classes that have been assigned to this synset do at least remotely fit the senses: more specific than `Impelling` and `Motion`, and equivalent to `Shooting`. Of course, a qualitative evaluation is needed to determine the degree of inaccuracy that is introduced. However, such an evaluation would rely heavily on manual inspection and could therefore not be carried out up to this moment.

4. Evaluation

4.1. Results and discussion

Table 1 below shows the results of the mapping procedure. Lines 1-3 in the table display the total number of synsets in the French EuroWordNet, as well as the numbers of those which have or have not received a SUMO mapping.

Deeper analysis of the 394 synsets which have not been assigned a SUMO concept reveals that almost 82% (323) represent terms from the new technology domain (line 12), such as computer terminology or internet vocabulary (e.g. *adresse d'inter-réseau* (*'network address'*), *applet* or *cache mémoire* (*'cache memory'*)). For 310 of these synsets, the reason for not receiving a SUMO concept is that although

[5]Possible ways of dealing with this issue will be discussed in Section 4.2. below.

	Type	Frequency	
		abs	rel
1	Synsets in French EuroWordNet	22,745	100.00%
2	... with SUMO mapping	22,351	98.27%
3	... without SUMO mapping	394	1.73%
	Of those with SUMO mapping		
4	... with one mapping	22,026	98.54%
5	... with two mappings	214	0.96%
6	... with three or more mappings	111	0.50%
7	... with only one sensemap	9,739	43.57%
8	... with more than one sensemap but only one SUMO class	12,287	54.97%
9	... with more than one sensemap and more than one SUMO class	325	1.46%
	Of those without SUMO mapping		
10	... nouns	324	82.23%
11	... verbs	70	17.77%
12	... from new technology domain	323	81.98%
13	... from food domain	8	2.03%
14	... collocational or idiomatic	23	5.84%

Table 1: Number of SUMO mappings grouped according to different types

Type	Frequency	
	abs	rel
SubjectiveAssessmentAttribute	1,293	5.78%
Device	1,088	4.87%
Artifact	689	3.08%
Motion	583	2.61%
OccupationalRole	555	2.48%
Communication	478	2.14%
Human	460	2.06%
Food	441	1.97%
SocialRole	404	1.81%
Process	379	1.70%
IntentionalProcess	361	1.62%
IntentionalPsychologicalProcess	276	1.23%
Text	247	1.11%
City	246	1.10%
StationaryArtifact	243	1.09%
NormativeAttribute	238	1.06%
EmotionalState	227	1.02%
Clothing	223	1.00%
DiseaseOrSyndrome	220	0.98%
FloweringPlant	205	0.92%

Table 2: Distribution of the top 20 assigned SUMO classes

they are part of the Inter-Lingual-Index, they were not part of Princeton WordNet 1.5. Therefore, they could not be included in the sensemap step from 1.5 to 1.6, and thus the assignment of a SUMO class failed. The remaining 13 synsets from these domain – in addition to a further two synsets from other domains – were not even part of the ILI, and thus represent a group of synsets that has been specifically created within the French EuroWordNet.

The rest of the synsets that have not been assigned to a SUMO concept (71) cannot be attributed to a single domain, though the food domain appears to be the strongest one (e.g. *petit four* (a specific kind of pastry) or *sauce au chocolat fondu* (a specific kind of chocolate sauce); see line 13). However, at least 23 synsets can be considered collocational or idiomatic in nature, e.g. *tenir compte de* (*'to account for'*), *vendre la mèche* (*'to reveal a secret'*; lit. *'to sell the fuse'*) or *saigner quelqu'un à blanc* (*'to exploit someone'*; lit. *'to bleed someone to white'*). In fact, these 23 synsets make up almost one third of the 70 verbs that have not been assigned a SUMO class, and a closer examination as well as their formalisation and integration into SUMO is certainly interesting and desirable.

Of the 22,351 synsets which have been assigned a SUMO class (cf. lines 4-6), 98.54% have been assigned to exactly one class, whereas 0.96% have been mapped to two and 0.50% to three or more SUMO classes. In line 8 we see that almost 55% of the synsets that have been assigned SUMO classes occurred in multiple sensemaps, but were all mapped onto synsets belonging to the same SUMO class, while only 1.46% where mapped onto two or more SUMO classes (cf. line 9). This means that only 1.46% are in principle able to cause "conceptual clashes" when retaining the strategy presented in Section 3.2. above. Table 2 displays the 20 most frequent SUMO classes that have been mapped to synsets in the French EuroWordNet. The frequency indicates the number of synsets which have been mapped directly onto the respective SUMO class, so no accumulation of frequency counts along the SUMO hierarchy was made, since that would – most probably – leave the top 20 slots in the table to the top 20 nodes in the hierarchy. A synset such as `00058624-n` (cf. examples above), which has been mapped onto three different SUMO classes, counts for each of these classes.

4.2. Improving the mapping

In Section 3.2., we briefly discussed issues that may arise from performing the mapping steps through all versions of WordNet up to the latest one, i.e. that the inaccuracies caused by synsets that have been split from one version to the next multiply. Since a lexical resource that makes use of the latest mappings between Princeton WordNet and SUMO is highly desirable, e.g. with respect to interoperability across WordNets for different languages, we will discuss below possible solutions to this issue in the order of manual labour required.

We have identified basically three ways of dealing with this problem. The first option is to accept the fact that the overall degree of inaccuracy will rise by applying the proposed methodology, and to carry out a qualitative evaluation of the resulting resource in order to estimate whether this degree is within a range that can be resolved with a reasonable amount of manual effort.

A second option is to apply the mapping iteratively only to those synsets which have not been split up to the latest version of WordNet, and then to use their latest mapping to SUMO. In other words, the result of this process would be a heterogeneous resource with synsets containing the latest mapping, in addition to synsets that have mappings based on the very first WordNet-SUMO mapping. Although this resource would certainly be more up to date than the one presented here, the usability of a heterogeneous resource is

highly questionable, especially if changes to SUMO have been made in the meantime which would render any resource inconsistent that uses certain concepts in coexistence. Therefore, it should be considered to update the mappings of the split synsets manually.
The last option – which would produce the resource with the highest quality, though at the cost of a very high amount of manual effort – is to create a direct mapping of the French EuroWordNet to the latest version of Princeton WordNet manually.
Orthogonal to these options, it should be considered to use the WordNet sense keys (Fellbaum, pers.comm.), which are available for later versions of WordNet and which would assure higher accuracy in the mapping in general.

5. Conclusion

We have presented a generic method for mapping EuroWordNets to the Suggested Upper Merged Ontology and have shown its application to the French EuroWordNet. The mapping procedure builds on existing work on SUMO and version 1.6 of Princeton WordNet (Niles and Pease, 2003), EuroWordNet's Inter-Lingual-Index (Vossen et al., 1998) and WordNet's sensemap files. The resulting resource largely conforms to the format suggested by the Global WordNet Grid initiative.
In the future, we intend to carry out a larger qualitative evaluation for the mapping procedure, which would investigate the accuracy of those synsets which – due to their changes from WordNet versions 1.5 to 1.6 – have been mapped onto more than one SUMO concept (cf. Section 3.2.). In addition to this, the ideas addressed in Section 4.2. will certainly be subject of future research. Moreover, we intend to use the resulting mapping to calculate ontological selectional preferences of French verbal predicates directly from corpora, with the ultimate goal to use these extracted selectional preferences for word sense disambiguation of the verbal predicates themselves as well as their arguments. A preliminary study has been presented in (Spohr, 2008), and we will intensify work on this in the near future.
Finally, it would be interesting to see the application of the mapping methodology to other EuroWordNets, provided that they are linked to the Inter-Lingual-Index as well. We do, however, expect our methodology to be generic enough to be applied to other languages without any major issues.

Acknowledgements

The research described in this work has been carried out as part of the project *'Polysemy in a Conceptual System'* (project B5 of SFB 732) and was funded by grants from the German Research Foundation. I should like to thank Christiane Fellbaum, Adam Pease, Achim Stein, and Piek Vossen for their valuable comments and suggestions.

6. References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In Christian Boitet and Pete Whitelock, editors, *Proceedings of ACL/COLING*, pages 86–90, San Francisco, California. Morgan Kaufmann Publishers.

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

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. The OntoWordNet project: extension and axiomatization of conceptual relations in WordNet. In Robert Meersman, Zahir Tari, and Douglas C. Schmidt, editors, *Lecture Notes in Computer Science*, number 2888, pages 820–828. Springer.

Michael R. Genesereth. 1991. Knowledge Interchange Format. In James Allen, Richard Fikes, and Erik Sandewall, editors, *Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning*, pages 238–249. Morgan Kaufmann Publishers.

Aleš Horák, Karel Pala, and Adam Rambousek. 2008. The Global WordNet Grid Software Design. In Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the 4th Global WordNet Conference*, pages 194–199, Szeged, Hungary.

Ian Niles and Adam Pease. 2001. Towards a Standard Upper Ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Ogunquit, ME.

Ian Niles and Adam Pease. 2003. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Hamid R. Arabnia, editor, *Proceedings of the IEEE International Conference on Information and Knowledge Engineering (IKE '03)*, pages 412–416, Las Vegas, NV.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

Nils Reiter. 2007. Towards a Linking of FrameNet and SUMO. Master's thesis, Saarland University.

Jan Scheffczyk, Adam Pease, and Michael Ellsworth. 2006. Linking FrameNet to the Suggested Upper Merged Ontology. In Brandon Bennett and Christiane Fellbaum, editors, *Proceedings of Formal Ontology in Information Systems (FOIS-2006)*, pages 289–300. IOS Press.

Dennis Spohr. 2008. Extraction of Selectional Preferences for French using a Mapping from EuroWordNet to the Suggested Upper Merged Ontology. In Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the 4th Global WordNet Conference*, pages 428–440, Szeged, Hungary.

Piek Vossen, Laura Bloksma, Horacio Rodriguez, Salvador Climent, Nicoletta Calzolari, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, and Wim Peters. 1998. *The EuroWordNet Base Concepts and Top Ontology*. Deliverable D017, D034 & D036.

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