Enriching Slovene WordNet with domain-specific terms

Špela Vintar, Darja Fišer
Dept. of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana, Slovenia
spela.vintar@guest.arnes.si, darja.fiser@guest.arnes.si

The paper describes an innovative approach to expanding the domain coverage of wordnet by exploiting multiple resources. In the experiment described here we are using a large monolingual Slovene corpus of texts from the domain of informatics to harvest terminology from, and a parallel English-Slovene corpus and an online dictionary as bilingual resources to facilitate the mapping of terms to the Slovene Wordnet. We first identify the core terms of the domain in English using the Princeton Wordnet, and then we translate them into Slovene using a bilingual lexicon produced from the parallel corpus. In the next step we extract multi-word terms from the Slovene domain-specific corpus using a hybrid approach, and finally match the term candidates to existing Wordnet synsets. The proposed method appears to be a successful way to improve the domain coverage of Wordnet as it yields abundant term candidates and exploits various multilingual resources.

Keywords: Wordnet construction, multi-word expressions, parallel corpora, term extraction, Slovene Wordnet.

1 Introduction

WordNet (Fellbaum 1998) is an extensive lexical database in which words are divided by part of speech and organized into a hierarchy of nodes. Each node represents a concept and words denoting the same concept are grouped into a synset with a unique id (e.g. ENG20-02853224-n: {car, auto, automobile, machine, motorcar}). Concepts are defined by a short gloss (e.g. 4-wheeled motor vehicle, usually propelled by an internal combustion engine) and are also linked to other relevant synsets in the database (e.g. hypernym: {motor vehicle, automotive vehicle}, hyponym: {cab, hack, taxi, taxicab}). Over time, WordNet has become one of the most valuable resources for a wide range of natural language processing applications, which initiated the development of wordnets for many other languages as well1.

One of such enterprises is the building of Slovene wordnet (Erjavec and Fišer 2006, Fišer 2007, Fišer and Sagot 2008). While this task would normally involve substantial manual labour and the efforts of several linguists, Slovene wordnet was built almost single-handedly exploiting multiple multilingual resources, such as bilingual dictionaries, parallel corpora and online semantic resources.

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

A combination of all these approaches yielded the first version of the Slovene wordnet[2] (sloWNet) containing about 17,000 synsets and 20,000 literals. However, the majority of these literals are single-word items, because the main lexicon extraction procedures involved in the building of WordNet involved no systematic handling of multi-word expressions. Also, the Slovene wordnet can only be as good as the resources that had been used for its construction. While the coverage for some domains, such as botany or zoology, is excellent, other domains remain underrepresented with numerous lexical gaps still to be filled. If we wish to use wordnet in any domain-specific application, such as Word Sense Disambiguation or Machine Translation, it is crucial that it contains the terminology of the target domain. The purpose of this paper is to propose a method to enrich wordnet with domain-specific single- and multi-word expressions.

The target domain in the experiments described below is information technology (IT), for which we have a 15 MW monolingual corpus and a small 300 kW parallel corpus. We use automatic term recognition to extract multi-word IT terms from the large Slovene corpus and word alignment to extract a bilingual lexicon of single-word terms from the parallel corpus. Using this lexicon and a domain-specific bilingual dictionary as a bridge across the two languages we connect the Slovene multi-word terms to the wordnet hierarchy via English, ie. the Princeton WordNet.

The rest of the paper is organized as follows: first, the Slovene WordNet Project is described. Section 3 describes the resources used and the procedure to extract domain-specific expressions from the corpus. Section 4 presents the bilingual part of the experiment where we try to map terms to the wordnet hierarchy. The results are discussed and evaluated in Section 5, and the paper ends with concluding thoughts and plans for future work.

2 Building the Slovene Wordnet

The first version of the Slovene wordnet was created on the basis of the Serbian wordnet (Krstev et al. 2004), which was translated into Slovene with a Serbian-Slovene dictionary. The main advantages of this approach were the direct mapping of the obtained synsets to wordnets in other languages and the density of the created network. The main disadvantage was the inadequate disambiguation of polysemous words, therefore requiring extensive manual editing of the results. The core Slovene wordnet contains 4,688 synsets, all from Base Concept Sets 1 and 2.

In the process of extending the core Slovene wordnet we tried to leverage the resources we had available, which are mainly corpora. Based on the assumption that translations are a plausible source of semantics we used multilingual parallel corpora such as the Multext-East (Erjavec and Ide 1998) and the JRC-Acquis corpus (Steinberger et al. 2006) to extract semantically relevant information (Fišer 2007).

[2] SloWNet is distributed under the Creative Commons licence, http://nl.ijs.si/slownet

We assumed that the multilingual alignment based approach can either convey sense distinctions of a polysemous source word or yield synonym sets based on the following criteria (cf. Dyvik 1998, Diab 2000 and Ide et al. 2000):

a) senses of ambiguous words in one language are often translated into distinct words in another language (e.g. Slovene equivalent for the English word *'school'* meaning educational institution is *'šola'* and *'jata'* for a large group of fish);
b) if two or more words are translated into the same word in another language, then they often share some element of meaning (e.g. the English word *'boy'* meaning a young male person can be translated into Slovene as either *'fant'* or *'deček'*).

In the experiment, corpora for up to five languages (English, Slovene, Czech, Bulgarian and Romanian) were word-aligned with Uplug (Tiedemann 2003) used to generate a multilingual lexicon that contained all translation variants found in the corpus. The lexicon was then compared to the existing wordnets in other languages. For English, the Princeton WordNet (Fellbaum 1998) was used while for Czech, Romanian and Bulgarian, wordnets developed in the BalkaNet project (Tufiş 2000) were used. If a match between the lexicon and wordnets across all the languages was found, the Slovene translation was assigned the appropriate synset id. In the end, all the Slovene words sharing the same synset ids were grouped into a synset.

The results obtained in the experiment were evaluated automatically against a manually created gold standard. A sample of the generated synsets was also checked by hand. The results were encouraging, especially for nouns with f-measure ranging between 69 and 81%, depending on the datasets and settings used in the experiment. However, the approach had two serious limitations: first, the automatically generated network contains gaps in the hierarchy where no match was found between the lexicon and the existing wordnets, and second, the alignment was limited to single-word literals, thus leaving out all the multi-word expressions.

We tried to overcome this shortcoming with extensive freely available multilingual resources, such as Wikipedia and Eurovoc. These resources are rich in specialized terms, most of which are multi-word. Since specialized terminology is typically monosemous, a bilingual approach sufficed to translate monosemous literals from PWN 2.0 into Slovene. A bilingual lexicon was extracted from Wikipedia, Wiktionary and Wikispecies by following inter-lingual links that relate two articles on the same topic in Slovene and English. We improved and extended this lexicon with a simple analysis of article bodies (capitalization, synonyms extraction, preliminary extraction of definitions). In addition we extracted a bilingual lexicon from Eurovoc, a multilingual thesaurus that is used for classification of EU documents. This procedure yielded 12.840 synsets. Translations of the monosemous literals are very accurate and include many multi-word expressions, and thus neatly complement the previous alignment approach. Also, they mostly contain specific, non-core vocabulary.

Synsets obtained from all three approaches were merged and filtered according to the reliability of the sources of translations. The structure of PWN synsets for which no translation could be found with any of the approaches was adopted from PWN based on the hierarchy preservation principle (Tufiş 2000), only the literals were left empty. The entre network of synsets was then formatted in DEBVisDic XML (Horak 2005). The latest version of sloWNet (2.1, 30/09/2009) contains about 20,000 unique literals, which are organized into almost 17,000 synsets, covering about 15% of PWN. Base Concept Sets 1 & 2 are fully covered but there are also many specific synsets. The most frequent domain in sloWNet is Factotum (25%) which was mostly obtained from the dictionary and a parallel corpus while the following three are Zoology (17%), Botany (13%) and Biology (7%) and come from Wikipedia.

sloWNet mostly contains nominal synsets (91%), and there are some verbal and adjectival synsets as well. Apart from single word literals, there are also quite a few multi-word expressions (43%). These too mostly come from Wikipedia. Synsets in sloWNet are relatively short as 66% of them contain only one literal, average synset length being 1.16 literals. The longest synset contains 16 literals (for verb *goljufati*, Eng. *to cheat*). The most common relation in sloWNet is hypernymy, which represents almost half of all relations in wordnet (46%). Hypernymy is by far the most prevalent relation for nouns (91%). Nominal hypernymy chains tend to be quite long, the longest ones containing 16 synsets. Since sloWNet does not cover the entire inventory of PWN concepts, there are some gaps (empty synsets) in the network. An investigation of nominal hierarchies revealed that almost half (46%) of the chains do not contain a single gap and that there are only 2% of chains with five or more gaps. These gaps will have to be filled in the future in order to obtain a denser hierarchy of nodes.

3 Harvesting domain-specific terminology from specialised corpora

3.1 Multi-word expressions and wordnet

Multi-word expressions (MWE) are lexical units that include a range of linguistic phenomena, such as nominal compounds (e.g. *blood vessel*), phrasal verbs (e.g. *put up*), adverbial and prepositional locutions (e.g. *on purpose, in front of*) and other institutionalized phrases (e.g. *de facto*). MWEs constitute a substantial part of the lexicon, since they express ideas and concepts that cannot be compressed into a single word. Moreover, they are frequently used to designate complex or novel concepts. As can be seen in Table 2, the majority of MWEs in Princeton Wordnet do not belong into any of the Basic Concept Sets, meaning that they encode specialized concepts and are frequently terms.

As a consequence, their inclusion into wordnet is of crucial importance, because any kind of semantic application without appropriate handling of MWEs is severely limited.

POS	Freq.
nouns	60,931
verbs	4,315
adverbs	955
adjectives	739
total	66,940

Table 1: The distribution of MWEs in PWN3 across part-of-speech

For the purpose of MWE identification, various syntactical (Bourigault 1993), statistical (Tomokiyo and Hurst 2003) and hybrid semantic-syntactic-statistical methodologies (Piao et al. 2003, Dias and Nunes 2004) have been proposed, to name but a few. Since the majority of MWEs included in the Princeton WordNet are nominal (see Table 1) and compositional, our approach is based on syntactic features of MWEs.

Group	Freq.
other	64,205
BCS 3	1,470
BCS 2	926
BCS 1	339
total	66,940

Table 2: The distribution of MWEs in PWN across BCS

In addressing the issue of MWEs in sloWNet, we initially wanted to find Slovene equivalents for the MWEs already present in Princeton Wordnet. We describe this experiment and its successful implementation in (Vintar and Fišer 2008).

3.2 Resources

If a wordnet is to be used in a semantic application within a specific domain, we wish to ensure its coverage within this domain primarily for the target language. The goal we address here is thus how to enrich sloWNet with domain-specific Slovene MWEs regardless of whether their English counterparts are included in PWN or not.

The resources we use to this end are the following (Figure 1):

- Ikorpus, a Slovene corpus of Computer Science texts, size ca. 15 million words, morphosyntactically annotated and lemmatized,
- a Slovene-English parallel corpus of Computer Science abstracts, size ca. 300,000 words, morphosyntactically annotated and lemmatized,
- Islovar, a Slovene-English online dictionary of Computer Science[4],

[3] The figures were taken from Princeton WordNet 2.1.

- Princeton WordNet.

The idea underlying our approach is that a large domain-specific corpus, especially one sufficiently varied in terms of register and text types, can be an excellent source of domain knowledge. Using terminology extraction, gloss extraction and relation extraction and mapping these to an existing semantic structure such as wordnet can help us construct a valuable domain-specific semantic resource for any language and with minimum manual effort. However, in order to map the extracted terms in the target language onto wordnet, we need a bilingual resource, preferably a domain-specific one, to provide the links between the source structure (in our case PWN) and the target structure (sloWNet). For our target domain of information science we have compiled a small parallel corpus of scientific abstracts and combined it with a bilingual online dictionary of computer science. Since both of these bilingual resources are used primarily to translate the hypernyms of the extracted terms, the parallel corpus does not need to fulfill all the requirements of a representative corpus.

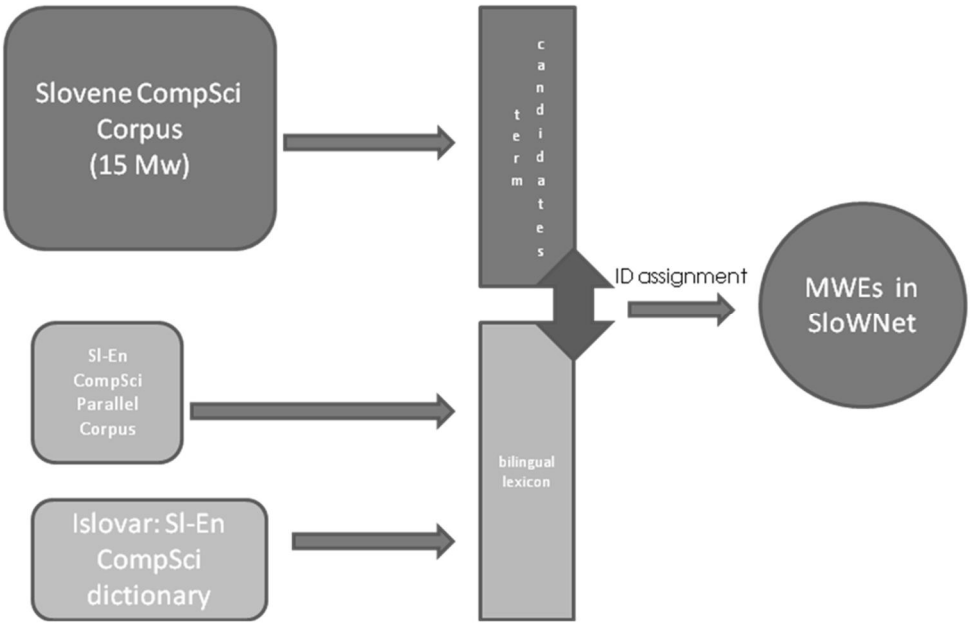

Figure 1: Resources for harvesting MWEs

3.3 Automatic Term Extraction

The domain-specific Ikorpus is composed of texts from 5 journals dealing with computer science, information and communication technology, and it also contains 5 consecutive volumes of proceedings of the largest informatics conference in Slovenia DSI. Its size is approximately 15 million running words, which makes it an excellent and fairly representative source of terminology.

[4] Islovar, a Slovene-English online dictionary of informatics, http://www.islovar.org

The task of automatically identifying domain-specific terms in texts has been addressed by numerous authors and has been tackled to the extent that there now exist several commercial tools with term extraction functionality. The main approaches described in literature range from statistical ones, where terms are viewed as kinds of collocations and the challenge lies in identifying the optimal word dependency measure (Dunning 1993; Daille 1995), to more linguistically informed and hybrid approaches where part-of-speech, morphology and syntax are exploited as indicators of termhood (Heid 1999; Dias et al. 2000). More recent approaches introduce semantics and utilize context features to detect terminologically relevant phrases in running text (Maynard and Ananiadou 1999; Gillam et al. 2007), as well as propose methods for the identification of term variants (Jacquemin 2001). An overview of the trends is given in Kageura et al. (2004).

In our experiment, automatic term extraction is performed using a hybrid approach based on morphosyntactic patterns for Slovene and statistical ranking of candidates (Vintar 2004, Vintar 2009). The patterns, such as Adjective+Noun or Noun+Noun[Gen], yield numerous potential MWEs. After candidate phrases had been extracted from the corpora, a term weighting measure is used to assign a "termhood" value to each phrase. The termhood value W of a candidate term a consisting of n words is computed as

$$W(a) = \frac{f_a^2}{n} * \sum_{1}^{n} (\log \frac{f_{n,D}}{N_D} - \log \frac{f_{n,R}}{N_R})$$

where f_a is the absolute frequency of the candidate term in the domain-specific corpus, $f_{n,D}$ and $f_{n,R}$ are the frequencies of each constituent word in the domain-specific and the general language reference corpus respectively and N_D and N_R are the sizes of these two corpora in tokens.

The rationale of the termhood measure is that terms are composed of terminologically relevant words, and the measure of terminological relevance is the comparison of a word's frequency between a domain-specific corpus and a general language corpus. This intuitive notion was first exploited by Ahmad et al. (1992) and implemented in other term extraction tasks (Scott 1998, Heid et al. 2001), however mostly for single-word terms. We use a modified version of this idea by adjusting it to multi-word expressions and including the frequency of the entire expression to override non-terminological phrases occurring only once.

For example, if we compare two phrases a and b, both occurring in the corpus of computer science texts, *spletni brskalnik [web browser]* (517) and *kakovost izdelka [product quality]* (74), using the 619-million-words FidaPlus corpus as the source of comparative frequencies we get the following result which indicates that the first phrase is terminologically more relevant than the second:

$$W(a) = 517^2/2 * (9{,}18 + 2{,}93) = 1618434{,}89$$

$$W(b) = 74^2/2 * (1.31 + 1.20) = 6872{,}38$$

The term extraction procedure performed on the 15-million-token Slovene corpus of computer science yielded over 70,000 term candidates of length up to 4 words (Table 3). Given this bulk we can safely assume that not all of them are really terms we would like to include in the Slovene wordnet. As it turns out, the candidates list contains a large number of named entities, such as names of software and hardware products, vendors and manufacturers. Since few of these names might be terminologically relevant, we excluded them from further processing. We also employed a frequency threshold and discarded all term candidates which occurred less than 5 times.

The extractor uses morphosyntactic patterns, therefore each multi-word term candidate, e.g. *domenski strežnik [domain name server]*, can be automatically assigned a headword (*strežnik [server]*) and we assume this to be the hypernym of the term candidate.

MWE size	**Number of candidates**
2 words (Adj+N, N+N, ...)	54,844
3 words (Adj + Adj + N, N + Prep + N, ...)	16,861
4 words (Adj + Adj + N + N, ...)	2,605
Total	**74,310**

Table 3: Term candidates and their length in words

Clearly, the domain-specific terms constitute a valuable lexical resource, but not until we can introduce some semantic structure. The next step therefore is to integrate at least some of these terms into the Slovene wordnet.

4 Mapping terms to Slovene wordnet

At this point we have a large number of Slovene multi-word terms without any semantic information other than the headword of each unit. Thus, for a term such as *prosto programje [free software]*, since it has been extracted through the syntactic pattern Adjective + Noun, we know that *programje* is the headword and *prosto* the modifier. We may also assume that *programje [software]* is the hypernym of *prosto programje [free software]*, and hence we could add *prosto programje [free software]* into Slovene wordnet as the hyponym of *programje [software]*, but only if the Slovene wordnet already contains the required headword *programje*.

For many multi-word terms this turns out not to be the case, which is why we wish to add both the hypernym and its extracted hyponyms to sloWNet in order to fill as many lexical gaps as possible. We use the Princeton Wordnet as the source of semantic structure, and to be able to link headwords to PWN we use bilingual lexicon extraction.

4.1 Bilingual lexicon extraction

Bilingual lexicon extraction, also known as word alignment, is a statistical procedure where for each source word *a* the algorithm computes the probabilities of all of its potential translation equivalents t_1, t_2, t_n in the target language (Och and Ney 2003). The translation equivalents with the highest probability scores are then proposed as entries in the bilingual lexicon. Bilingual lexicon extraction can only be performed on parallel corpora or bitexts.

A small English-Slovene parallel corpus of 300,000 tokens was fed to the Uplug word aligner (Tiedemann 2003), which produced suggested translations for each word found in the corpus. To improve accuracy, we use only alignments of words that occur more than once and alignment scores over 0.05. This yields a bilingual single-word lexicon of 1326 words, mostly nouns (Table 4).

Freq	Score	English	POS	Slovene	POS
4	0.058264988	adaptability	n	prilagodljivost	n
8	0.100445189	additional	a	dodaten	a
5	0.138443460	agent	n	agent	n

Table 4: Sample entries in the bilingual lexicon

In order to improve coverage and accuracy, the automatically extracted bilingual lexicon was further enlarged with entries from the English-Slovene online dictionary of computer science Islovar. The dictionary provided approximately 5,000 bilingual entries and was consulted also in certain cases of ambiguous headword, as described below.

4.2 Adding Terms to sloWNet

For each Slovene multi-word term candidate we first identify its headword and assume that the headword is its hypernym. Using our bilingual lexicon we translate the headword into English and retrieve its synset IDs from PWN. If the headword turns out to be monosemous, the entire group of multi-word terms with the same hypernym can be added to the Slovene wordnet under the unique synset ID (Table 5).

If the headword could be assigned several possible senses, we exploit the domain label in wordnet, such as *factotum, biology* etc. If one of the senses of the polysemous headword belongs to the domain Computer Science, than this sense is chosen (Table 6).

Term candidates	**Hypernym, English translation and possible synset IDs**	**Selected synset ID**
prosto programje [free software] *priloženo programje [attached software]* *ustrezno programje [appropriate software]* *novejše programje [updated software]* *dodatno programje [additional software]* *vohunsko programje [spyware]*	programje = **software** **ENG20-06162514-n** [computer_science]	**ENG20-06162514-n**

Table 5: Monosemous headword

If the headword is already part of the Slovene wordnet, no disambiguation is needed and the terms can be simply added as hyponyms to the existing Slovene hypernym. Also, in some cases one of the extracted multi-word terms was already in the Islovar dictionary. We can then use the English translation of the term to look up the correct hypernym and synset ID in PWN.

Nevertheless there remain many cases where the polysemous headword does not belong to the CompSci domain in wordnet and it is neither included in wordnet or Islovar. In such cases the correct sense must be picked manually (Table 7).

Term candidates	**Hypernym, English translation and possible synset IDs**	**Selected synset ID**
vgrajena tipkovnica [built-in keyboard] *brezžična tipkovnica [wireless keyboard]* *zaslonska tipkovnica [monitor keyboard]* *tipkovnica qwerty [QWERTY keyboard* *navidezna tipkovnica [virtual keyboard]* *miniaturna tipkovnica [miniature keyboard]* *zunanja tipkovnica [external keyboard]* *zložljiva tipkovnica [folding keyboard]* *ergonomska tipkovnica [ergonomic keyboard]* *programska tipkovnica [program keyboard]* *slovenska tipkovnica [Slovene keyboard]* *modularna tipkovnica [modular keyboard]* *alfanumerična tipkovnica [alphanumeric keyboard]*	tipkovnica = **keyboard** ENG20-03480332-n [computer_science] ENG20-03480198-n [factotum]	ENG20-03480332-n

Table 6: Polysemous headword with CompSci domain

Term candidates	Hypernym, English translation and possible synset IDs	Selected synset ID
nalaganje gonilnikov [loading drivers] *nalaganje podatkov [loading data]* *nalaganje programov [software download]* *nalaganje strani [loading page]*	nalaganje = **loading** ENG20-00671518-n [factotum] ENG20-13044298-n [transport]	to be selected manually

Table 7: Polysemous headword, ID to be selected manually

5 Discussion

Extracting terms from a large domain-specific Slovene corpus yielded the bulk of 74,310 term candidates. We keep only those that occur more than 5 times and where the headword and its English translation can be identified with reasonable accuracy, and we disregard all names and terms that include names. Some of the remaining terms were already either in the Islovar dictionary or in sloWNet, however the large majority were new. Table 8 shows the number of terms successfully added to sloWNet.

Category	Number of terms
Already in sloWNet	29
Already in PWN	23
Already in Islovar	198
New	5150
Total	**5400**

Table 8: Total term candidates added to sloWNet

The assumption that the headword of the multi-word expression is at the same time the hypernym of the term may seem daring, however we encountered very few examples where this is not the case. Within a random sample of 200 multi-word terms we found 5 terms where the headword could not be considered an appropriate hypernym of the term, for example a *spletni portal [web portal]* is not a kind of *portal [portal]*; *portal* being an architectural term, and *prostor na disku [disk space]* is not a kind of *prostor [space]*; although both of these headwords could be used elliptically in a computer science context to mean *[web portal]* or *[disk space]* respectively.

As has been described in the previous section, the difficult part is determining the correct sense of the potentially polysemous headword. This ambiguity can of course affect a large number of terms, since – as can be seen in Table 6 – several dozens of multi-word terms share the same headword. While we use all the semantic information we can infer either from the domain label or the online dictionary, nearly half of all the headwords need to be disambiguated manually (Table 9).

Category	**Number of headwords**
Monosemous	84
Headwords with CompSci domain	35
Headwords already in sloWNet	11
Headwords derived from MWE PWN	6
To be picked manually	136
Total	**272**

Table 9: Categories of headwords

In this respect our methodology could benefit significantly from additional context-based disambiguation procedures. A possible approach would be to use the contexts of the polysemous headwords and compute the semantic similarity between the relevant context words and each sense of the headword. The sense with the greatest semantic similarity to the context features is selected as the correct one. This is essentially a word disambiguation task and various authors have proposed similarity measures based on the graph representation of wordnet (e.g. Leacock and Chodorow 1998; Wu and Palmer 1994; Agirre et al. 2009). In future experiments we plan to implement such methods for the selection of the correct sense.

Finally it should be noted that the domain labels in Princeton Wordnet are sometimes illogical, too specific or not specific enough. If we for example explore the financial domain, there is no label [finance], but we find three different domains for a related set of concepts: *money* [money], *coin* [money], *bank* [banking], *account* [banking], *pay* [economy]. This is clearly a problem for automatic text processing, because we cannot rely on the fact that semantically related lexical items share the same domain label in wordnet. On the other hand there exists a hierarchical structure of wordnet domains which was not taken into account in our experiments. It may be the case that some ambiguity issues could be better resolved using this hierarchy.

6 Conclusions

We described an approach to improve the domain coverage of wordnet by enriching it with semi-automatically extracted multi-word terms. Our method utilizes a combination of mono- and bilingual resources. A large monolingual domain-specific corpus is used as the source of terminology, and a smaller parallel corpus combined with a domain-specific dictionary is used to provide translation equivalents of headwords. These are required in order to map the semantic structure of Princeton Wordnet onto the Slovene term candidates and thus integrate them into sloWNet.

Although the approach works well and yields many items of specialised vocabulary, the most difficult part is the selection of the correct sense with polysemous headwords. In some cases the correct sense can be inferred from the domain label or from the dictionary, but in many cases this step still has to be

performed manually. In the future we plan to implement a sense disambiguation procedure based on semantic similarity.

It should be noted that an evaluation of monolingual term extraction lies beyond the scope of this paper and is not addressed, although the quality of the term candidates clearly influences the results of the experiment described. Term extraction evaluation depends heavily on the target application, which means that the same system may perform very well in an information retrieval task and poorly in a dictionary-making task. Since the measure of terminological relevance relies on the comparison of relative frequencies between a domain-specific and a reference corpus, the term extraction system performs better for highly specialised domains or, in other words, for terms that do not occur frequently in general language. Information science is in this respect not the ideal domain because IT-related topics are regularly discussed in general language media.

The proposed methodology can be extended to other domains, or indeed other languages. While we employ a specialised monolingual corpus, a bilingual corpus and a specialised bilingual dictionary, the cross-language part of the algorithm is essentially suited to parallel corpora. Especially in domains – or language pairs – for which bilingual dictionaries are scarce it is often more viable to construct a small parallel corpus and use the word-aligned bilingual lexicon to translate headwords. While in other domains we could again exploit the domain labels in wordnet to disambiguate the headword, our methodology is less suitable for general language where polysemy is common and disambiguation can only be performed with context-based methods.

An evaluation of the domain coverage of sloWNet will be performed within a Machine Translation application. In the future we also plan to extend this approach to the extraction of definitions from domain-specific corpora using Machine Learning to distinguish between well-formed and not-well-formed definitions (Fišer et al. 2010).

7 References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, A., Pasca, M. & Soroa, A. 2009. "A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches." *Proceedings of NAACL-HLT 09*. Boulder, USA.

Ahmad, K., Davies, A., Fulford, H. & Rogers, M. 1992. "What is a term? The semi-automatic extraction of terms from text." In Snell-Hornby et al. (Eds.), *Translation Studies – an interdiscipline*. Amsterdam/Philadelphia: John Benjamins.

Bourigault, D. 1993. "Analyse syntaxique locale pour le repérage de termes complexes dans un texte." *Traitement Automatique des Langues*, 34 (2), 105-117.

Daille, B. 1995. "Combined Approach For Terminology Extraction: Lexical Statistics And Linguistic Filtering". *COLING 94*, 515-521.

Diab, M. & Resnik, P. 2002, "An Unsupervised Method for Word Sense Tagging using Parallel Corpora". *40th Anniversary Meeting of the Association for Computational Linguistics* (ACL-02), Philadelphia, July, 2002.

Dias, G., Guilloré, S., Bassano, J. C., Lopes, J. G. P. 2000. "Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association?". *Proceedings of 6ème Conférence sur la Recherche d'Informations Assistée par Ordinateur (RIAO 2000)*, Paris, France, 1-20.

Dias, G. & Nunes, S. 2004. "Evaluation of Different Similarity Measures for the Extraction of Multiword Units in a Reinforcement Learning Environment". *Proceedings of the 4th International Conference On Languages Resources and Evaluation*, Lisbon, Portugal, 1717-1721.

Dunning, T. 1993. "Accurate methods for the statistics of surprise and coincidence". *Computational linguistics*, 19, 61-74.

Dyvik, H. 1998. "Translations as semantic mirrors." *Proceedings of Workshop W13: Multilinguality in the lexicon II*, Brighton, UK. The 13th biennial European Conference on Artificial Intelligence ECAI 98, 24-44.

Erjavec, T. & Fišer, D. 2006. "Building Slovene WordNet". *Proceedings of the 5th International Conference on Language Resources and Evaluation LREC'06*, Genoa, Italy.

Erjavec, T. & Ide, N. 1998. "The MULTEXT-East Corpus". In *Proceedings of the First International Conference on Language Resources and Evaluation, LREC'98*, Granada, Spain.

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

Fišer, D. & Sagot, B. 2008. "Combining multiple resources to build reliable wordnets". In *Text, Speech and Dialogue Conference (LNCS 2546)*. Berlin/Heidelberg: Springer, 61-68.

Fišer, D. 2007. "Leveraging Parallel Corpora and Existing Wordnets for Automatic Construction of the Slovene Wordnet". *Proceedings of the 3rd Language and Technology Conference L&TC'07*, Poznan, Poland.

Fišer, D., Pollak, S. & Vintar, Š. 2010. "Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources". *Proceedings of the 7th Language Resources and Evaluation Conference*, Malta.

Gillam, L., Tariq, M. & Khurshid, A. 2007. "Terminology and the construction of ontology". In F. Ibekwe-SanJuan et al. (Eds.), *Application-Driven Terminology Engineering*. Amsterdam/Philadelphia: John Benjamins, 49-74

Heid, U. 1999. "Extracting Terminologically Relevant Collocations from German Technical Texts". In P. Sandrini (Ed.), *Terminology and Knowledge Engineering (TKE99)*. Vienna: TermNet, 241-255.

Heid, U., Evert, S., Fitschen, A., Freese, M. & Vögele, A. 2001. *Term candidate extraction in DOT. Dot final report, Part II*. Stuttgart: IMS, University of Stuttgart.

Horak, A., Pala, K., Rambousek, A. & Povolny, M. 2005. "DEBVisDic - First Version of New Client-Server Wordnet Browsing and Editing Tool". *Proceedings of the Global Wordnet Conference GWA'05*, Brno, 325–328.

Ide, N., Erjavec, T. & Tufis, D. 2002. "Sense Discrimination with Parallel Corpora". *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, 54-60.

Jacquemin, C. 2001. *Spotting and Discovering Terms through Natural Language Processing*. Cambridge, Massachussetts: MIT Press.

Kageura, K., Daille, B., Nakagawa, H. & Chien, L.-F. 2004. "Recent trends in computational terminology." *Terminology*, 10 (1), 1-21.

Krstev, C., Pavlović-Lažetić, G., Vitas, D. & Obradović, I. 2004. "Using textual resources in developing Serbian wordnet." *Romanian Journal of Information Science and Technology*, 7 (1-2), 147-161.

Leacock, C. & Chodorow, M. 1998. "Combining local context and WordNet sense similarity for word sense identification". In *WordNet, An Electronic Lexical Database*. The MIT Press.

Maynard, D. & Ananiadou, S. 1999. "Term Extraction Using a Similarity-Based Approach". In *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins.

Och, F. J. & Ney, H. 2003. "A Systematic Comparison of Various Statistical Alignment Models". *Computational Linguistics*, 29 (1), 19-51.

Piao, S., Rayson, P., Archer, D., Wilson A. & McEnery, T. 2003. "Extracting Multiword Expressions with a Semantic Tagger". *Workshop on Multiword Expressions of the 41st ACL meeting*, Sapporo, Japan, 49-57.

Scott, M. 1998. "Focusing on the Text and Its Key Words". *TALC 98 Proceedings*, Oxford, 152-164.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. & Varga, D. 2006. "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages". *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

Tiedemann, J. 2003. *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*, Doctoral Thesis. Studia Linguistica Upsaliensia 1.

Tomokiyo, T. & Hurst, M. 2003. "A Language Model Approach to Keyphrase Extraction". *Workshop on Multiword Expressions of the 41st ACL meeting*, Sapporo, Japan, 33-41.

Tufis, D. 2000. "BalkaNet - Design and Development of a Multilingual Balkan WordNet". *Romanian Journal of Information Science and Technology Special Issue*, 7 (1-2).

Vintar, Š. 2004. "Comparative Evaluation of C-value in the Treatment of Nested Terms". *Memura 2004 - Methodologies and Evaluation of Multiword Units in Real-World Applications (LREC 2004)*, 54-57.

Vintar, Š. & Fišer, D. 2008. "Harvesting Multi-Word Expressions from Parallel Corpora". *Proceedings of the 6th International Conference on Language Resources and Evaluation LREC'08*, Marrakech, Morocco.

Vintar, Š., 2009. "Samodejno luščenje terminologije - izkušnje in perspective". [Automatic Term Recognition – Experience and Perspectives]. In N. Ledinek, M. Žagar Karer, M. Humar (Eds.), *Terminologija in sodobna terminografija*. Ljubljana: Založba ZRC, 345-356.

Wu, Z. & Palmer, M. 1994. "Verb semantics and lexical selection". *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico.