Harvesting Multi-Word Expressions from Parallel Corpora

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

Abstract

The paper presents a set of approaches to extend the automatically created Slovene wordnet with nominal multi-word expressions. In the first approach multi-word expressions from Princeton WordNet are translated with a technique that is based on word-alignment and lexico-syntactic patterns. This is followed by extracting new terms from a monolingual corpus using keywordness ranking and contextual patterns. Finally, the multi-word expressions are assigned a hypernym and added to our wordnet. Manual evaluation and comparison of the results shows that the translation approach is the most straightforward and accurate. However, it is successfully complemented by the two monolingual approaches which are able to identify more term candidates in the corpus that would otherwise go unnoticed. Some weaknesses of the proposed wordnet extension techniques are also addressed.

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 NLP applications, which initiated the development of wordnets for many other languages as well[1].

One of such enterprises is the building of Slovene wordnet (Erjavec and Fišer 2006). The approach is based on extracting translation equivalents from parallel corpora with automatic word-alignment and disambiguating the polysemous words with wordnets that already exist for other languages. This paper is an extension of our previous experiment and tries to overcome the limitation of the alignment approach with a complementary method to add multi-word expressions harvested from the JRC-Acquis parallel corpus (Steinberger et al. 2006) to Slovene 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 a consequence, their inclusion into wordnet is of crucial importance, because any kind of semantic application without appropriate handling of MWEs is severely limited.

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 below) and compositional, our approach is based on syntactic features of MWEs.

The rest of the paper is organized as follows: first, the Slovene WordNet Project is described. Section 3 describes the procedure used to extract multi-word expressions from the corpus and their mapping to the wordnet hierarchy. The results are presented and evaluated in Section 4, and the paper ends with concluding thoughts and plans for future work.

2. The Slovene WordNet Project

The first version of the Slovene wordnet was created on the basis of the Serbian wordnet (Krstev et al. 2004) 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).

[1] See http://www.globalwordnet.org/gwa/wordnet_table.htm [15.03.2008]

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.

3. Adding Multi-Word Expressions to Slovene Wordnet

This section presents the three approaches we used to add multi-word expressions to Slovene wordnet. The first approach is bilingual and uses a lexicon we extracted from the automatically word-aligned JRC-Acquis corpus and lexico-syntactic patterns to find Slovene translations of English multi-word expressions that are in Princeton WordNet (PWN).

The second two approaches are monolingual and aim to complement the translation approach by identifying additional multi-word expressions that are not yet in PWN. Domain-specific MWEs were extracted from a fishing subcorpus of the JRC-Acquis and filtered using fishing-related seedwords found in Eurovoc[2].

In the final approach, we use lexico-syntactic patterns to extract semantically related expressions (hypernyms and hyponyms) from the corpus and then map them to Slovene wordnet.

3.1 Translation-Based Approach: Adding MWEs Found in The Princeton WordNet

First, a list of multi-word expressions PWN was extracted. As Table 1 shows, most of them were nouns. Since our alignment-based approach proved to work best on nouns, we decided to limit our work to those (about 61,000). A large majority of the expressions do not belong to Base Concept Sets (about 64,000) that are the main constituents of the current version of Slovene wordnet. This means that this approach will generate a lot of new synsets, extending our wordnet even further (see Table 2).

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

Table 1. The distribution of MWEs in PWN across POS

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

Table 3 shows that the polysemy of MWEs is not nearly as high as polysemy of single-word literals: a large majority of expressions were found in only one synset and very few were highly polysemous. This is why we were able to restrict the number of languages used in the experiment to English and Slovene only.

no. of synsets	no. of MWE's
1	56,989
2	3,335
3	453
4	108
5	40
6	6
total	60931

Table 3. Polysemy of MWEs in PWN

Our first task was to match English MWEs found in PWN with their Slovene counterparts in the JRC-Acquis corpus. The chosen approach combines methods of bilingual lexicon extraction (Tiedeman 2003) with those of terminology extraction (e.g. Bourigault et al. 1996, Martin and Heid 2001).

[2] http://europa.eu/eurovoc/ [15.03.2008]

For each source MWE from PWN, we extracted all sentence pairs from the parallel corpus that contain the source term. Also, for each single word from the source MWE we extract all possible translation equivalents from the bilingual lexicon. We then use lexico-grammatical patterns to identify potential multi-word terms in the target language and employ the bag-of-equivalents approach for the selection of the best equivalent, which is the candidate with the most matches for each constituent word in the bilingual lexicon (Vintar 2004).

For each source MWE this approach yields a list of target MWE's, ranked according to the quality of single-word translations inherited from the word-alignment system. Adding the translated MWE's to the Slovene wordnet is trivial in this case because each English MWE already has a unique synset id from PWN that is retained in Slovene wordnet (e.g MWE *'art exhibition'* with PWN id 'ENG20-07896855-n' is translated into Slovene as *'umetnostna razstava'* and keeps the same id in Slovene wordnet).

3.2 Seed-Word-Based Approach: Extracting Domain-Specific MWEs

With the second approach we tried to identify completely new MWEs that were not found in PWN and add them to Slovene wordnet. Synsets that were added in this way were assigned PWN-like synset ids that retain the source information (e.g. ENG20-08419438-n vs. JRC01-00000001-n) Our monolingual approach was tested on the domain of fishing. A subcorpus of nearly 2,000 documents was created from the Slovene part of the JRC-Acquis.

From this corpus we extracted multi-word terms using a hybrid method based on part-of-speech patterns and keywordness ranking. The patterns included noun phrases such as A N N (adjective + noun + noun) or A N S N (adjective + noun + preposition + noun). The extracted noun phrases were then ranked according to their keywordness, where each constituent word of a MWE is assigned its keywordness by comparing its relative frequency in the fishing subcorpus to its relative frequency in the FidaPlus corpus, a 600-million-word reference corpus of Slovene (Arhar et al. 2007).

This yields quite a large number of terms related to fishing, but since we needed to establish a link to existing synsets in the Slovene Wordnet, we selected a small number of fishing-related seed words from the multilingual EUROVOC descriptors *(riba, ribolov, ribištvo, mreža, ladja, ladjevje, plovilo [fish, fishing, fishery, net, ship, fleet, vessel])*. It was now possible to select all extracted MWEs that contained one of the seed words. If the seed word was the headword of the extracted MWE (e.g. *vlečna mreža [trawl net] – hyponym of – mreža [net]*), we presuppose the MWE to be a hyponym of the seed word, whereas if the seed word appears in a pre- or postmodifying position, the hypernym of the MWE is its headword (e.g. *posadka plovila [vessel crew] – hyponym of – posadka [crew]*).

3.3 Pattern-Based Approach: Extracting New Semantically Related MWEs

In our final approach, we extracted knowledge-rich contexts from the entire JRC-Acquis corpus using lexico-grammatical patterns (Finkelstein-Landau and Morin 1999; Malaisé et al. 2007), e.g.:

[NP] , kot so ([NP],)+ [NP] in|ali [NP]
([NP] , such as ([NP],)+ [NP] and|or [NP])

[NP] , kot na primer ([NP],)+ [NP] in|ali [NP]
([NP] , for example ([NP],)+ [NP] and|or [NP])

([NP],)+ [NP] in|ali drug [NP]
([NP],)+ [NP] and|or other [NP])

Such contexts typically contain a set of semantically related terms, usually with one term representing a superordinate (parent) concept (e.g. *regional fishing organisation*) and one or several other terms representing subordinate (child) concepts (e.g. *NAFO, ICCAT, CCAMLR, NEAFC*). Since no parser for Slovene is publicly available at the time, we used a self-made shallow parser to identify noun phrases. The JRC-Acquis corpus contains over 12,000 'kot so (such as)' contexts, however not all yield structures suitable for inclusion into Slovene wordnet. The challenge in extracting relevant candidates for the taxonomic is-a relation lies especially in accurate identification of constituent terms.

Far from trivial is also the task of finding the corresponding wordnet synset into which the terms extracted from the corpus should be included. This was done in several steps:

- First, we checked whether the hypernym was already in Slovene Wordnet. In this case, its hyponyms extracted from the corpus were just added to it.
- If the hypernym was not found in wordnet, it was decomposed step by step until a match was found. For example, the term *'mednarodna pomorska organizacija' [international naval organization]* was first decomposed to *'pomorska organizacija' [naval organization].* Because there was no match, the term was further decomposed to *'organizacija' [organization]* and this expression was then considered as the hypernym of our term. If there were more possible hypernyms for a given term, the term was added to all hypernym candidates with a mark requiring manual inspection.
- If the hypernym could not be found in Slovene Wordnet at all, we checked whether any hyponyms were already in wordnet. The rest of the extracted cohyponyms were treated as its siblings and attached to the same parent node.

The difference between this approach and the approach described in Section 3.2 is that we could not filter the contexts with any domain-sensitive filters in order to reduce noise. As it turns out, a lot of the extracted contexts were useless for our purpose because the semantic relations between the words were highly contextual and therefore not suitable for inclusion to the taxonomy (e.g. *postavke, kot so status, privilegij in imuniteta [items, such as status, privilege and immunity]*). Also, many hypernyms used were too general to be of any informative value in taxonomy population (e.g. *dejavniki, kot so poreklo, spol in starost [factors, such as origin, sex and age]).*

4. Evaluation and Discussion of the Results

In this section we present the results of the experiment, which are evaluated manually. The evaluation consisted of checking whether a given MWE candidate is in fact an appropriate term and whether it was mapped to Slovene wordnet correctly. Each approach was first evaluated separately, then the findings were combined and compared.

4.1 Translation-Based Approach

The goal of this experiment was to find Slovene equivalents for MWEs found in the Princeton WordNet using the English-Slovene part of the JRC-Acquis corpus (see Section 3.1). Since the equivalents proposed by the algorithm are ranked according to the average scores of their constituent single-word alignments, we set a threshold of 0.05 as the lowest possible similarity score. In this way we obtained a list of Slovene MWEs which were then checked by hand. Manual evaluation shows that 85% of MWEs were correctly translated into Slovene. These (1,059) were then all added to Slovene wordnet by retaining synset ids attached to each MWE we found in PWN (see Table 4).

English MWE	**Slovene MWE**	**Equivalence**
benzoic acid	benzojska kislina	0.1551
economic and social council	ekonomsko-socialni svet	0.1529
photographic camera	fotografska kamera	0.1516
republic of mozambique	republika mozambik	0.1485
tensile strength	natezna trdnost	0.1452
red meat	rdeče meso	0.1448
brassica rapa	brassica rapa	0.1437

Table 4: Example of translated MWEs

The advantage of this approach is that it is quite accurate, but also that it will extract less common translations for a source term as well (e.g. *'milking cow'* is correctly translated as *'krava mlekarica'* as well as *'krava dojilja'*). The quality of the results is however highly dependent on the quality of the word alignment, and – in turn – lemmatisation and POS tagging. Most errors that occurred during translation of an English MWE into Slovene were due to incomplete lexico-syntactic patterns for Slovene, thus extracting incomplete noun phrases. There were no errors in mapping synsets to Slovene wordnet.

4.2 Seed-Word-Based Approach

In this approach, we first extracted a list of multi-word expressions from the fishing subcorpus, which was then filtered using a list of seed terms taken from Eurovoc (see Section 3.2). Here, all the expressions were extracted correctly and are indeed fishing-related terms, because the restriction was that they must contain at least one seed word. We obtained however only 251 terms, 75 of which contain the seed word as the headword and thus directly imply their hyponymic relation to it, while 176 candidates contain the seed word either as a premodifier or postmodifier.

Hypernym assignment for terms that contain a seedword was almost perfect. All but 2 headwords were correctly identified in Slovene wordnet and the terms were added to them as their hyponyms. The two headwords that were missing in the wordnet *('plovilo' [vessel]* and *'mreža' [net]*) were compensated by finding one of their hyponyms in the network ('ribiško *plovilo' [fishing vessel]* and *'ribiška mreža' [fishing net]*) and treating the rest of the terms as their cohyponyms.

The approach was slightly less successful in cases where seedwords were pre- or postmodifiers of MWEs. Wordnet contained hypernyms for 126 out of such 165 MWEs and 5 of the 175 hyponyms. However, in about a third of the cases there were more possible hypernym candidates. Since no automatic WSD was available, the correct hypernym was therefore selected by hand.

While our approach presupposes that MWE's are compositional, i.e. that a MWE's headword is at the same time its hypernym, this is not always the case, as the examples below show:

morski meč – *hyponym of – meč
[sea sword (silver scabbardfish)] - [sword]

morski pes – *hyponym of – pes
[sea dog (shark)] – [dog]

It turned out that only 6 out of 175 MWEs were non-compositional. For these the appropriate hypernym could not be identified from their constituent words and could therefore not be correctly added to wordnet.

4.3 Pattern-Based Approach

In the last approach we used lexico-syntactic patterns to extract words from the corpus that are in a hypernym-hyponym relation. The extracted candidates were then mapped to wordnet (see Section 3.3). Table 5 gives the results of candidate extraction and mapping. It must be noted here that not all NP's we harvested from the corpus were MWE's and are therefore not included in evaluation and discussion in this paper.

Pattern	***such as***	***for example***	***and other***
Extracted	941	38	3344
Mapped	503	29	1817
MWE	217	14	1385

Table 5: Yield of lexico-grammatical patterns

The biggest drawback of this approach was the nature of semantic relatedness of words matching the pattern, which is highly contextual in the JRC-Acquis corpus and too vague to be included in the wordnet. Also, many hypernyms that were successfully identified in both the extracted terms and wordnet are too general to be of any real taxonomic significance. This is why many (almost half) of the extracted patterns were not mapped to Slovene wordnet at all.

4.4 Comparison of the Approaches

As table 6 shows, by far the most MWEs were obtained from the pattern approach. But very few of them were (correctly) added to wordnet. On the other hand, the seedword approach generated relatively few MWE candidates but a large majority of those were successfully mapped to wordnet. The best approach is clearly the first one in which all of the correctly translated MWEs were also successfully added to wordnet. This is hardly surprising because the English MWEs were exctracted from PWN and therefore already had appropriate synset ids which were simply inherited by their Slovene translation equivalents. This approach was also attractive because it simply reused the existing resources that were created in our previous experiments.

The value of the second two approaches lies in the fact that there are many more useful MWEs in corpora than in PWN. These cannot be handled by the first approach that relies on PWN, which is why complementary monolingual techniques can be handy in identification of additional MWEs, thus generating more synset candidates for the extension of the wordnet.

	translation approach	**seed-word approach**	**pattern approach**	**total**
correctly extracted	1,059	251	4,323	5,597
correctly added to wordnet	1,059 (100%)	206 (82%)	1616 (37%)	2,881 (52%)

Table 6: Comparison of the approaches according to the number of correctly extracted MWEs and the number of correctly added MWEs to Slovene wordnet

5. Conclusion

The paper has proposed several methods to incorporate multi-word expressions in a lexico-semantic resource, such as wordnet. Manual evaluation of the results obtained in the experiment to enrich Slovene wordnet with MWEs shows that the translation-based approach works best. However, monolingual seedword-based and pattern-based approaches give additional synset candidates that the first one cannot detect. With a combination of all three approaches described in the paper we were able to add almost 3000 MWEss to Slovene wordnet.

In the future we wish to improve our purely regular expression-based NP identifier, which outputs noun phrases that do match the pattern but are not really terms and should therefore be improved.

Another necessary improvement of the procedure is the automation of the selection of the best hypernym candidate that is currently being done by hand, which is tedious and slow.

The problem that many semantic relations extracted from the corpus are contextual rather than taxonomic remains unresolved here. Also, the density of the hierarchy remains an issue and will have to be dealt with in our future work. A possible solution would be to use more encyclopaedic-like resources such as Wikipedia in which the semantic relations between words are more regular.

6. References

Arhar, Špela, Vojko Gorjanc & Simon Krek (2007): FidaPLUS corpus of Slovenian. The New Generation of the Slovenian Reference Corpus: Its Design and Tools. In *Proceedings of the Corpus Linguistics Conference* (CL2007), Birmingham, July, 2007.

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

Bourigault, Didier, Gonzales-Mullier, I. & Gros, C. (1996): LEXTER, a Natural Language Processing Tool for Terminology Extraction. In: Gellerstam, M. et al. (eds.): Euralex '96 Proceedings I-II. Göteburg: Universität Göteburg, 771--780.

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

Dias, Gael & Nunes, S. (2004): Evaluation of Different Similarity Measures for the Extraction of Multiword Units in a Reinforcement Learning Environment. In M.T. Lino, M.F. Xavier, F. Pereira, R. Costa and R. Silva (eds): Proceedings of the 4th International Conference On Languages Resources and Evaluation, M.T. Lino, M.F. Xavier, F. Pereira, R. Costa and R. Silva (eds), Lisbon, Portugal, May 26-28. 1717--1721.

Dyvik, Helge (1998): Translations as semantic mirrors. In *Proceedings of Workshop W13: Multilinguality in the lexicon II*, pp. 24.44, Brighton, UK. The 13th biennial European Conference on Artificial Intelligence ECAI 98.

Erjavec, Tomaž & Darja Fišer (2006): Building Slovene WordNet. In: Proceedings of the 5th International Conference on Language Resources and Evaluation LREC'06. 24-26th May 2006, Genoa, Italy.

Erjavec, Tomaž & Ide, Nancy (1998): The MULTEXT-East Corpus. In: Proceedings of the First International Conference on Language Resources and Evaluation, LREC'98. Granada, Spain.

Fellbaum, Christiane (1998): WordNet: An Electronic Lexical Database. MIT Press.

Finkelstein-Landau, M. & E. Morin (1999): Extracting semantic relationships between terms: Supervised vs. unsupervised methods. In International Workshop on Ontological Engineering on the Global Information Infrastructure, pages 71--80.

Ide, Nancy, Tomaž Erjavec & Dan Tufis (2002): Sense Discrimination with Parallel Corpora. In *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, 54--60.

Krstev, Cvetana, G. Pavlović-Lažetić, D. Vitas & I. Obradović (2004): Using textual resources in developing Serbian wordnet. In: Romanian Journal of Information Science and Technology. (Volume 7, No. 1-2), pp 147--161.

Malaisé, Véronique, Pierre Zweigenbaum & Bruno Bachimont (2007): Mining defining contexts to help structuring differential ontologies. In: Ibekwe-SanJuan et al. Application-Driven Terminology Engineering. Amsterdam: John Benjamins, pp.19--47.

Piao, S., Rayson, P., Archer, D., Wilson, A. & McEnery, T. (2003): Extracting Multiword Expressions with a Semantic Tagger. In Workshop on Multiword Expressions of the 41st ACL meeting. 7-12 July. Sapporo. Japan. 49--57.

Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş & Dániel Varga (2006): The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy, 24--26 May 2006.

Tiedemann, Jörg (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. In Workshop on Multiword Expressions of the 41st ACL meeting. 7-12 July. Sapporo. Japan. 33--41.

Tufis, Dan (2000): BalkaNet - Design and Development of a Multilingual Balkan WordNet. In: Romanian Journal of Information Science and Technology Special Issue (Volume 7, No. 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), pp. 54--57.