sloWNet 3.0: development, extension and cleaning

Darja Fišer
University of Ljubljana
Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Jernej Novak
University of Maribor
Maribor, Slovenia
jernej.novak1@uni-mb.si

Tomaž Erjavec
Jožef Stefan Institute
Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract

In this paper we present the development, extension and cleaning of Slovene wordnet by reusing existing language resources. The initial induction of synsets and the subsequent extension of sloWNet are based on multilingual resources and were performed automatically. The cleaning of the developed lexicon, on the other hand, is based on a monolingual reference corpus and requires manual validation. Manual work is performed in sloWTool, a new browser, editor and visualizer of wordnet content. The developed wordnet and editor are freely available under the Creative Commons licence.

1 Introduction

In the past five years much effort has been invested to close the gap in Slovene language resources which had still been lacking in the lexico-semantic layer. Semantic lexica and semantically annotated corpora are a prerequisite for practically any task that involves semantically enhanced processing of natural language. In addition, they are also extremely useful in applied linguistics, such as lexicography and language pedagogy, as well as in corpus linguistics for the study of sense frequency and co-occurrence.

While the development of the resources is still on-going and a lot of work on further extensions and refinements still needs to be done in the future, we have reached a relatively stable version of the resources which are large and precise enough to be useful in practical applications. The Slovene wordnet called sloWNet has been used to mine definitions from corpora (Fišer et al. 2010) improve the results of machine translation at lexical level (Fišer and Vintar 2010), to automatically detect semantic shifts in translated texts (Vintar 2011), and as a seed dictionary for building context vectors to extract bilingual lexicons from large comparable corpora (Ljubešić and Fišer, submitted).

The aim of this paper is to present the latest developments connected with the Slovene wordnet and is organized as follows: in the next section we summarize the development stages of sloWNet and report on the content of the latest version. In Section 3 we present the corpus that was annotated with wordnet synsets and the refinements that were made in order to achieve greater consistency. In Section 4 we describe the tool we developed for browsing, editing and visualizing wordnet content, and in Section 5 conclude with some final remarks and plans for future work.

2 Automatic development of sloWNet

Slovene wordnet is based on Princeton WordNet (Fellbaum 1998) and was built automatically in three stages, each using a different approach according to the resources used for extracting the relevant lexico-semantic information. The first and most straightforward approach relied on the Serbian wordnet (Krstev et al. 2004) where the literals were translated into Slovene utilizing a traditional digitized bilingual Slovene-Serbian dictionary (Erjavec and Fišer 2006). This simple approach lacked automatic disambiguation of polysemous dictionary entries and therefore required a lot of manual cleaning. This was improved in the second approach which was able to assign the correct wordnet sense to a Slovene equivalent by disambiguating it with a word-aligned parallel multilingual corpus and already existing wordnets for several languages (Fišer 2007). The main contribution of the third and final approach was the extraction of a large number of monosemous specialized vocabulary and

multi-word expressions from Wikipedia and its related resources (Fišer and Sagot 2008). After merging the results of these three approaches sloWNet contained about 17,000 literal which belonged to roughly 20,000 synsets.

3 Further extension of sloWNet

The next major step in the development of sloWNet 3.0 is the recent large-scale automatic extension in which we combined all the resources from the previous steps in order to exploit the available resources to their full potential and thereby improve coverage of sloWNet without compromising its quality. First, a model was trained on the existing elements in sloWNet, and a maximum entropy classifier was used to determine appropriate senses of translation candidates extracted from the heterogeneous resources described above (see Sagot and Fišer 2012).

The extended sloWNet now has 114% more synsets than before, while the number of *(literal, synset)* pairs has increased from 24,081 to 82,721, which is 244% of its initial size. An analysis of the sources that contributed to creation of the current synsets in sloWNet shows that for adjectives, the classifier provided as much as 95% of the synsets. Similarly, 88% of the adverbial synsets were populated by the classifier, while the rest were added by hand during manual revision of the created wordnet (see Secton 4). Verbs behave in a very similar fashion with 85% of them originating from the classifier while the rest originated in the previous development step from the dictionary, manual revision and the corpus. This is understandable because these parts of speech had been handled less successfully in our previous wordnet development approaches. But a somewhat different distribution can be seen among nominal synsets that had dominated the Slovene wordnet all along. Of these, 36% had already been created before from Wikipedia, dictionary and the corpus or were added manually, so that only 64% were contributed in the recent extension process by the classifier.

As Table 1 shows, the current version of Slovene wordnet contains 36% of all the synsets in Princeton WordNet. Nouns are still by far the most frequent, representing more than 70% of all synsets. sloWNet contains all synsets from the Base Concept Sets but also has a lot of specialized vocabulary as 66% of all the literals in it are monosemous. The extended sloWNet also contains a lot of multi-word expressions and proper names, which are both mostly nominal. A comparison of the average number of literals per synset and average level of polysemy between sloWNet and PWN is interesting because it can indicate how accurate the automatic population of Slovene synsets was. While average synset length is comparable to PWN, the total average polysemy (2.07 vs. 1.51) and the average polysemy excluding monosemous words (4.19 vs. 3.39) show that Slovene wordnet contains noise that will have to be filtered out in the future (see Section 4).

The fact that sloWNet is noisy due to the automatic construction process is further indicated by the number of literals in the longest synsets which are, at first glance, quite similar to PWN (see Table 2) but a more careful analysis shows that even though these synsets contain several synonyms, not all of them are correct and should therefore be filtered out in the future.

This is even more obvious with the most polysemous literals in sloWNet that are clearly very noisy. The most important source of such errors was the inadequate sense assignment for the most frequent words in the language, such as the verb *to be*, the noun *person*, the adjective *big* and the adverb *very*.

While Princeton WordNet contains glosses for all its 117,658 synsets, sloWNet currently contains only 3,178 definitions for nominal synsets that were extracted automatically from Wikipedia articles. 32,881 PWN synsets are also equipped with at least one usage example which is only the case for the 517 sloWNet nominal synsets that were annotated in the corpus. A focused attempt to providing additional definition and example sentences is planned in the near future.

As additional information, useful in for many tasks, many wordnet synsets have domain labels which are further organized into a domain hierarchy (Bentivogli et al. 2004). Analysis of the domains in the created Slovene wordnet shows that they are much better represented. 46% of all the synsets in PWN that belong to one of the domains exist in sloWNet as well. Of all 161 domains that are present in PWN, only 4 of them are missing entirely, all of them belonging to the Sports domain hierarchy: Rugby, Soccer, Sub and Volleyball. Just like in PWN, the most frequent domain is Factotum and the following three most frequent ones are represented in the same order in both wordnets. There are also many similarities among the ten most frequent domains in the two wordnets (see Table 3).

no. of synsets			no. of literals			no. of (synset, literal) pairs		
	PWN3.0	sloWNet3.0		PWN3.0	sloWNet3.0		PWN3.0	sloWNet3.0
Adj	18,156	6,218	Adj	21,538	5,108	Adj	30,004	12,438
Adv	3,621	453	Adv	4,481	514	Adv	5,580	847
N	82,114	30,911	N	119,034	30,319	N	146,345	55,383
V	13,767	5,337	V	11,531	3,840	V	25,047	14,053
total:	117,658	42,919	total:	156,584	39,781	total:	206,976	82,721
BCS1	1,220	1,220	monos.	130,208	26,339	avg. syn. length	1.76	1.92
BCS2	2,213	2,213	mwe	64,383	9,050	avg. polys.-all	1.51	2.07
BCS3	1,238	1,238	proper n.	35,002	2,946	avg. polys.-poly	3.39	4.19
total:	4,671	4,671	non-letter lit.	178	32			

Table 1: A comparison of Princeton WordNet 3.0 and sloWNet 3.0

longest synsets		
POS	PWN 3.0	sloWNet 3.0
Adj	23 (02074929-a)	23 (00148078-a)
Adv	10 (00048739-b)	14 (00004722-b)
N	28 (05559256-n)	20 (05921123-n)
V	25 (01426397-v)	24 (00933821-v)
most polysemous literals		
POS	PWN 3.0	sloWNet 3.0
Adj	27 (heavy)	47 (velik~big)
Adv	13 (well)	13 (zelo~very)
N	33 (head)	70 (oseba~person)
V	59 (break)	757 (biti~to be)

Table 2: A comparison of longest synsets and most polysemous literals in PWN 3.0 and sloWNet 3.0

PWN 3.0	Synsets	sloWNet 3.0	Synsets
Factotum	19,454	*Factotum*	9,701
Zoology	6,270	*Zoology*	3,345
Botany	5,998	*Botany*	2,716
Biology	3,004	*Biology*	1,512
Gastronomy	2,183	Person	793
Chemistry	2,011	Admin.	790
Medicine	1,999	Chemistry	656
Admin.	1,909	Medicine	625
Anatomy	1,768	Building_ind.	575
Person	1,600	Gastronomy	525
Total	77,701	total	33,126

Table 3: A comparison of synsets belonging to domains in PWN 3.0 and sloWNet 3.0

4 Cleaning of sloWNet synsets

Even before the large-scale automatic extension, sloWNet has undergone two cycles of partial manual revision; the goal of the first revision was to manually check, correct and add the missing translations for all synsets belonging to the Base Concept Sets (about 5,000) while the second revision was performed in parallel with semantic annotation of the corpus (see Fišer and Erjavec 2010). At that time, all the synsets containing the nouns (about 1,000, not all of which were finally assigned to words in the corpus) which were selected for annotation in the corpus were checked and corrected as necessary and the missing senses of those nouns were also added if they were found in the corpus.

But as the analysis of the automatically extended wordnet in the previous section shows, a more comprehensive cleaning of the resource is required. We have developed a 2-step procedure where we first automatically detected those literals that are most likely outliers given the synsets they appear in, which were then presented to wordnet editors for manual validation.

The automatic detection of noisy literals is based on distributional methods and aims to identify the most obvious errors in synsets that occurred due to errors in word-alignment of parallel corpora (e.g. misaligned elements of multiword expressions) and inappropriate word-sense disambiguation of homonymous words (e.g. assigning a valid translation of one sense of a homonymous source word to all its senses). We started from a (noisy) list of synonym candidates and ranked them according to the similarity of contexts they appear in FidaPLUS reference corpus (Arhar and Gorjanc 2007). The ranking relied on a simple hypothesis that literal-synset pairs tend to co-occur in corpora with other lexemes that are semantically related, as made explicit by relations between synsets in wordnet (see Sagot and Fišer, submitted).

So far, the procedure has been applied on nominal synsets only because they represent the majority (67%) of all synset-literal pairs in the latest version of sloWNet. Of all nominal synset-literal pairs in sloWNet, 37,356 (67%) were attested in the reference corpus. We then empirically set a threshold that defines the minimum score under which a (literal, synset) pair is considered as a candidate outlier. The overall error rate in the extended sloWNet has been evaluated at 15%, which means that around 13,000 incorrect (literal, synset) pairs were introduced with the automatic extension. We have therefore chosen a threshold for candidate outliers of the same order of magnitude, generating 12,578 candidate outliers. They are not deleted from wordnet automatically because some candidates could not be ranked reliably due to lacking distributional information in the corpus, which is why they are presented to wordnet editors who decide whether they are true outliers or not. Despite the manual effort required in the cleaning of the created wordnet, the approach is still valuable because instead of having to check all (literal, synset) pairs in sloWNet, the editors now check only about a third of them, saving them a significant amount of time and effort. With the experience gained during manual validation of the candidate outliers we plan to further refine their automatic detection as well as to extend the approach to other parts of speech.

The detected candidate outliers are currently being manually validated in sloWTool, an all-in-one browser, editor and visualizer for wordnets we developed for this and many other tasks (Fišer and Novak 2011). An example of this procedure is illustrated in Figure 1 where the polysemous English word *organ* has two possible translations in Slovene: *organ* for the body part sense, and *orgle* for the instrument sense. In the example shown, *orgle* is the correct literal for the synset but *organ* is not, which was correctly detected as an outlier candidate and therefore has to be manually deleted from the synset by clicking on the delete button right of the highlighted literal.

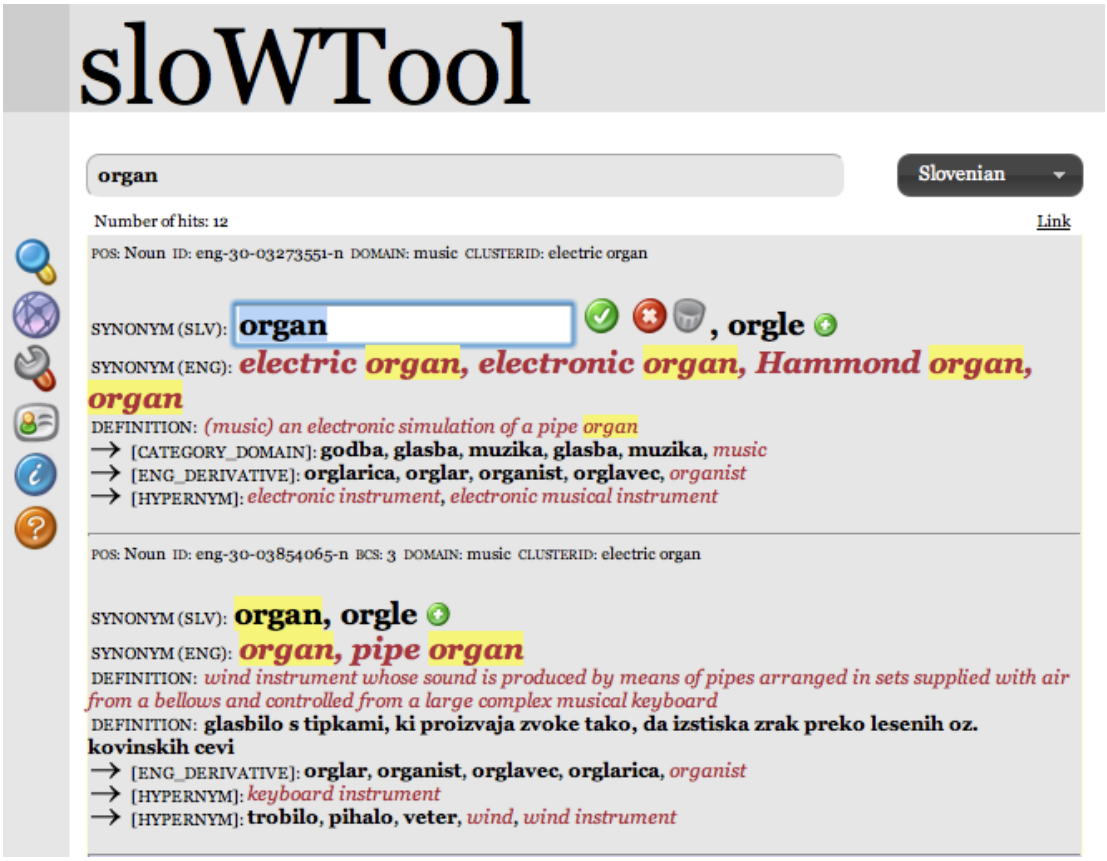

Figure 1: Manual editing of synsets in sloWTool.

5 Conclusions

In this paper we presented a set of automatic approaches we used to develop Slovene wordnet by recycling already available language resources. The analysis of the created sloWNet showed that filtering of inappropriate synset elements was required in order to reduce the noise in the resource, which is being done in a 2-step way by first automatically detecting candidate outliers based on distributional methods and then manual validation of the noisy candidates. Manual work is being performed in sloWTool, the 3-in-1 tool for wordnet browsing editing and visualisation. Both sloWNet and slowTool are freely available for research under the Creative Commons license at http://nl.ijs.si/slownet/.

In the future we plan to refine the approach to removing noisy synset candidates so that it can handle more subtle polysemy as well and to extend it to other parts of speech. We also wish to focus on refinements and extensions of the sloWNet-annotated jos100k corpus (Erjavec et al. 2010) so that it can serve as a training set for automatic word-sense disambiguation. In addition, we are looking into possibilities to further develop sloWTool to allow assigning wordnet senses to words in the corpus.

References

Arhar, Špela and Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa (The FidaPLUS corpus: a new generation of the Slovene reference corpus). *Jezik in slovstvo*, 52(2).

Luisa Bentivogli, Pamela Forner, Bernardo Magnini and Emanuele Pianta. 2004. Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing, *Proc. of the Workshop on Multilingual Linguistic Resources*, COLING'04, Geneva, Switzerland, August 28, 2004, pp. 101-108.

Tomaž Erjavec and Darja Fišer. 2006. Building the Slovene Wordnet: first steps, first problems. *Proc. of the Third International WordNet Conference (GWA'06)*, Jeju Island, Korea, January 22-26, 2006.

Tomaž Erjavec, Darja Fišer, Simon Krek and Nina Ledinek. 2010. The JOS linguistically tagged corpus of Slovene. *Proc. of 7th International Conference on Language Resources and Evaluation* (LREC'10), Malta, May 17-23.

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

Darja Fišer. 2007. Leveraging parallel corpora and existing wordnets for automatic construction of the Slovene wordnet. *Proc. of the 3rd Language & Technology Conference*, October 5-7, 2007, Poznań, Poland, pp. 162-166.

Darja Fišer and Benoît Sagot. 2008. Combining multiple resources to build reliable wordnets. Text, Speech and Dialogue (LNCS 2546). Berlin; Heidelberg: Springer, 2008 pp. 61-68.

Darja Fišer and Špela Vintar. 2010. Uporaba wordneta za boljše razdvoumljanje pri strojnem prevajanju. *Proc. of the 13th International Multiconference Information Society* (IS'10).

Darja Fišer, Senja Pollak and Špela Vintar. 2010. Learning to mine definitions from Slovene structured and unstructured knowledge-rich resources. *Proc. of 7th International Conference on Language Resources and Evaluation* (LREC'10), Malta, May 17-23 2010.

Darja Fišer and Jernej Novak. 2011. Visualizing sloWNet. Proc. of Electronic lexicography in the 21st century: new applications for new users (eLex'11), Bled, 10-12 November 2011.

Cvetana. Krstev, Gordana Pavlović-Lažetić, Duško Vitas and Ivan Obradović. 2004. Using textual resources in developing Serbian wordnet. *Romanian Journal of Information Science and Technology*. 7/1–2, pp 147-161.

Nikola Ljubešić and Darja Fišer, submitted. Extracting translation equivalents for polysemous words from comparable corpora.

Benoît Sagot and Darja Fišer, 2012. Automatic Extension of WOLF. Proc. of the 6th International Global Wordnet Conference (GWC'12), Matsue, Japan-January, 9-13, 2012.

Benoît Sagot and Darja Fišer, submitted. Cleaning Noisy Wordnets.

Špela Vintar. 2011. Samodejno odkrivanje semantičnih premikov v prevodih. Proc. of 30th simposium Obdobja, Ljubljana, Slovenia, November 19-19, 2011.