Combining multiple resources to build reliable wordnets

Darja Fišer,, Benoît Sagot

Combining multiple resources to build reliable wordnets

Darja Fišer[1], Benoît Sagot[2]
1. Fac. of Arts, Univ. of Ljubljana, Aškerčeva 2, 1000 Ljubljana, Slovenia
2. Alpage, INRIA / Paris 7, 30 rue du Ch. des rentiers, 75013 Paris, France
darja.fiser@guest.arnes.si, benoit.sagot@inria.fr

Abstract. This paper compares automatically generated sets of synonyms in French and Slovene wordnets with respect to the resources used in the construction process. Polysemous words were disambiguated via a five-language word-alignment of the SEERA.NET parallel corpus, a subcorpus of the JRC Acquis. The extracted multilingual lexicon was disambiguated with the existing wordnets for these languages. On the other hand, a bilingual approach sufficed to acquire equivalents for monosemous words. Bilingual lexicons were extracted from different resources, including Wikipedia, Wiktionary and EUROVOC thesaurus. A representative sample of the generated synsets was evaluated against the gold-standards.

1 Introduction

The first wordnet was developed for English at Princeton University (PWN). Over time it has become one of the most valuable resources in applications for natural language understanding and interpretation, which initiated the development of wordnets for many other languages apart from English [1, 2]. Currently, wordnets for more than 50 languages are registered with the Global WordNet Association (`http://www.globalwordnet.org/`). While it is true that manual construction of each wordnet is the most reliable and produces the best results as far as linguistic soundness and accuracy is concerned, such an endeavour is highly time-consuming and expensive. This is why alternative, semi- or fully automatic approaches have been proposed. By taking advantage of the existing resources, they facilitate faster and easier development of a wordnet [3, 4].

Apart from the knowledge acquisition bottleneck, another major problem in the wordnet community is the availability of the developed wordnets. Currently, only a handful of them are freely available (Arabic, Hebrew, Irish and Princeton). For example, a wordnet for French has been created within the EuroWordNet (EWN) project [1], the resource has not been widely used mainly due to licensing issues. In addition, there has been no follow-up project to further extend and improve the core French WordNet since the EWN project has ended [5]. This issue was taken into account in the two recent wordnet development projects presented in this paper, the results of which will be automatically constructed (but later also manually checked) broad-coverage open-source wordnets for French (WOLF, `wolf.gforge.inria.fr`) and Slovene (SloWNet, `nl.ijs.si/slownet`).

The paper is organized as follows: a brief overview of the related work is given in the next section. Section 3 presents the two wordnet development projects. Section 4 presents and evaluates the created resources with a focus on a source-by-source evaluation, and the last section gives conclusions and perspectives.

2 Related work

The relevant literature reports on several techniques used to build semantic lexicons, most of which can be divided into two approaches. Contrary to the merge approach, according to which a wordnet for a certain language is first created based on monolingual resources and then mapped to other wordnets, we have opted for the expand approach [1]. This model takes a fixed set of synsets from Princeton WordNet (PWN) and translates them into the target language, preserving the structure of the original wordnet. The cost of the expand model is that the resulting wordnets are biased by the PWN. However, due to its greater simplicity, the expand model has been adopted in a number of projects, such as the BalkaNet [2] and MultiWordNet [6], as well as EWN [1].

Research teams adopting the latter approach took advantage of a wide range of resources at their disposal, including machine readable bilingual and monolingual dictionaries, taxonomies, ontologies and others. For the construction of WOLF and SloWNet, we have leveraged three different publicly available types of resources: the JRC-Acquis parallel corpus, Wikipedia (and other related wiki resources) and other types of bilingual resources. Equivalents for monosemous literals that do not require sense disambiguation were extracted from bilingual resources. Roughly 82% of literals found in PWN are monosemous, however most of them are not in the core vocabulary. On the other hand, the parallel corpus was used to obtain semantically relevant information from translations so as to be able to handle polysemous literals. The idea that semantic insights can be derived from the translation relation has been explored by [7–9]. The approach has also yielded promising results in an earlier smaller-scale experiment to obtain synsets for Slovene wordnet [10].

3 Approach

This section briefly presents the approach used to construct a wordnet automatically. For a more detailed description of the approach, see [11].

In the align approach we used the SEE-ERA.NET corpus (project ICT 10503 RP), a 1.5-million-word sentence-aligned subcorpus of JRC-Acquis [12] in eight languages. Apart from French and Slovene, we used English, Romanian, Czech and Bulgarian. We used different tools to POS-tag and lemmatize the corpus before word-aligning it with Uplug [13]. This allowed us to build five multilingual lexicons that include French and four multilingual lexicons that include Slovene. They contain between 49,356 (Fr-Ro-Cz-Bg-En) to 59,020 entries (Fr-Cz-Bg-En). The next step was to assign the appropriate synset id to each entry of these lexicons. To achieve this, we gathered the set of all synset ids assigned

to each literal of a given entry (apart from the French or Slovene one) in the corresponding BalkaNet wordnet [2]. Since all these wordnets share the same synset ids as PWN 2.0, the intersection of all the found synset ids is computed. The intersection of all possible senses in each language is likely to output the correct one, which can be assigned to the French or Slovene literal. Applied to the above-mentioned multilingual lexicons, this technique allowed us to build several sets of (French or Slovene) synsets (see Table 2 for quantitative data). Because tagging, lemmatization and alignment are not perfect, synsets created in this way do inherit some of these errors. However, the value of this approach lies in the fact that they cover polysemous literals from the core vocabulary, which the translation approach cannot handle (see Section 4).

For the translation approach, applied on monosemous literals from the PWN 2.0, we used the following bilingual resources:

- Wikipedia (`http://www.wikipedia.org`), a multilingual collaborative encyclopaedia. We extracted bilingual Fr-En and Sl-En lexicons thanks to interwiki links that relate articles on the same topic in different languages.[1]
- The French, Slovene and English Wiktionaries (`http://www.wiktionary.org`), lexical companions to Wikipedia, which contain translations into other languages.
- The Wikispecies (`http://species.wikimedia.org`), a taxonomy of living species with translations of Latin standard names into vernacular languages.
- Eurovoc descriptors (`http://europa.eu/eurovoc`) is a multilingual thesaurus used for classification of EU documents.
- For Slovene, we used a large-coverage electronic bilingual (English-Slovene) dictionary (over 130,000 entries).
- Finally, we created trivial translations by retaining all numeric literals (such as 1,000 or 3.14159...) and all Latin taxonomic terms (extracted from the TreeOfLife project — `www.tolweb.org`).

Because they are created by translation of monosemous literals, these synsets will on the one hand be very reliable (see Table 3), but at the same time mostly concern non-core vocabulary (see Table 1).

Synsets obtained from both approaches were merged, while preserving information on the source of each piece of information. This enabled us to perform a simple heuristic filtering according to the reliability of each source, on the diversity of sources that assign a given literal to a given synset, and on frequency information (for the sources from the align approach).

[1] These lexicons have 307,256 entries for French and 27,667 for Slovene. The difference in size is substantial and will also lead to very different number of the generated synsets. The same is true for most other bilingual resources used in this approach.

4 Results and evaluation

4.1 Global evaluation

We compared the merged Slovene and French wordnets to PWN, French EWN and a manually created sample of Slovene WordNet, called ManSloWNet[2]. Although we are aware of the fact that these resources are not perfect, they were considered as gold standard for our evaluation procedure because they were by far the best resources of such kind we could obtain.

WOLF currently contains 32,351 synsets that include 38,001 unique literals. This figure is much greater than the number of synsets in French EWN (22,121 synsets could be mapped into PWN 2.0 synsets). This is directly related to the high number of monosemous PWN literals in non-core synsets (119,528 out of 145,627) that the translation approach was able to handle adequately. Moreover, French EWN has only nominal and verbal synsets, whereas WOLF includes adjectival and adverbial synsets as well. Figures for SloWNet are similar: 29,108 synsets that include 45,694 literals (to be compared with the 4,868 synsets of ManSloWNet). However, without the En-Sl dictionary that was used for Slovene, the figures would have been much lower.

In order to evaluate the coverage of the generated wordnets, we used the BalkaNet *Basic Concept Sets* [2]. Basic synsets are grouped into three BCS categories, BCS1 being the most fundamental set of senses. The results for the automatically constructed wordnets are compared to the goldstandards (see Table 1). They show that both WOLF and SloWNet have a reasonable coverage of BCS senses. They also show that our approach still does not come close to PWN, which was the backbone of our experiment. However, the generated wordnets are considerably richer than the only other wordnets that exist for French and Slovene, especially for non-BCS synsets. Moreover, although the same approach was used, and despite the use of a bilingual dictionary, SloWNet is smaller than WOLF. This is mainly because French Wikipedia is considerably larger than the Slovene one and thus yields many more monosemous synsets, which are not always found in the En-Sl bilingual dictionary.

The align approach yielded a relatively low number of synsets compared to bilingual resources, mostly because it relies on an intersection operation among several languages: if some synsets were missing in any of the existing wordnets used for comparison, there was no match among the languages and the synset could not be generated. Interesting as well is the nature of synsets that were generated from the different sources. Basically, the align approach that handled all kinds of words resulted predominantly in core synsets from the BCS categories. On the other hand, the bilingual resources that tackled only the monosemous expressions provided us with much more specific synsets outside the core vocabulary. The align approach worked only on single words, which is why all MWEs in the resulting wordnets come from bilingual resources.

[2] This subset of the Slovene WordNet contains all synsets from BCS1 and 2 (approx. 5,000), which were automatically translated from Serbian, its closest relative in the BalkaNet family. All the synsets were then manually corrected [14].

Automatically generated French synsets (WOLF)

wordnet	PWN 2.0	align	transl	merged (WOLF)	French EWN
BCS1	1,218	791 64.9%	175 14.4%	**870 71.4%**	1,211 99.4%
BCS2	3,471	1,309 37.7%	523 15.1%	**1,668 48.0%**	3,022 87.1%
BCS3	3,827	824 21.5%	1,100 28.7%	**1,801 47.1%**	2,304 60.2%
non-BCS	106,908	2,844 2.7%	25,566 23.9%	**28,012 26.2%**	15,584 14.6%
total	*115,424*	*5,768 5.0%*	*27,364 23.7%*	***32,351 28.0%***	*22,121 19.2%*

Automatically generated Slovene synsets (SloWNet)

wordnet	PWN 2.0	align	transl	merged (SloWNet)	ManSloWNet
BCS1	1,218	618 50.7%	181 14.9%	**714 58.6%**	1,218 100%
BCS2	3,471	896 25.8%	606 17.4%	**1,361 39.2%**	3,469 99.9%
BCS3	3,827	577 15.1%	1,128 29.5%	**1,611 42.1%**	180 4.7%
non-BCS	106,908	1,603 1.5%	24,116 22.6%	**25,422 23.8%**	1 0.0%
total	*115,424*	*3,694 3.2%*	*26,031 22.6%*	***29,108 25.2%***	*4,868 4.2%*

Table 1. WOLF and SloWNet synsets. Percentages are given compared to PWN 2.0.

4.2 Source-by-source evaluation

From a qualitative point of view, we were interested in how reliable the various sources, such as the different language combinations in the align approach, wiki sources and other thesauri, were for the creation of synsets. This is why the rest of the evaluation is performed on each individual source and the reliability scores obtained will be used to generate confidence measures for the rest of the generated synsets from that source which we were unable to evaluate automatically (because they are missing in the goldstandards). We restricted this manual evaluation to nominal synsets. Verbal synsets are more difficult to handle automatically for many reasons: higher polysemy of frequent verbs, differences in linguistics systems in dealing with phrasal verbs, light verb constructions and others. These synsets, as well as adjectival and adverbial synsets, will be evaluated carefully in the future. For each source, we checked whether a given literal in the generated wordnets is assigned the appropriate synset id according to the goldstandards. We considered only those literals that are both in the goldstandard and in the evaluated resource. A random sample of 100 (literal,synset) pairs present in the acquired resource but absent in the goldstandard were inspected by hand and classified into the following categories (see Table 3):

- the literal is an appropriate expression of the concept represented by that synset id but is missing from the goldstandard (absent in GS but correct); as mentioned before, the goldstandards we used for automatic evaluation are not perfect and complete, which is why a given literal that was automatically assigned to a particular synset can be a legitimate literal missing in the goldstandard rather than an error; for example, the French literal *document* and the Slovene literal *dokument* were correctly added in the synset corresponding to PWN literal *document*: this synset was absent from French EWN altogether, whereas in ManSloWNet it only contained literal *spis*;

Source	# of (lit,synsetid) pairs	Present in GS	synset not in GS	Discrepancy w.r.t GS
WOLF				
Fr-Cz-En	1760	61.7%	7.5%	30.8%
Fr-Cz-Bg-En	1092	67.8%	4.9%	27.4%
Fr-Ro-En	2002	64.7%	8.1%	27.2%
Fr-Ro-Cz-En	1206	70.6%	5.4%	24.0%
Fr-Ro-Cz-Bg-En	796	75.5%	3.3%	21.2%
Wikipedia	368	94.0%	0.3%	5.7%
Fr Wiktionary	577	69.8%	1.0%	29.1%
En Wiktionary	365	88.5%	-	11.5%
Wikispecies	21	*90.5%*	*4.8%*	*4.8%*
EUROVOC descr.	69	67.6%	-	32.3%
SloWNet				
Sl-Cz-En	2084	53.4%	10.9%	35.6%
Sl-Cz-Bg-En	1383	59.3%	6.6%	34.1%
Sl-Ro-Cz-En	1589	57.7%	8.0%	34.3%
Sl-Ro-Cz-Bg-En	1101	61.0%	5.1%	33.9%

Table 2. Evaluation of WOLF and SloWNet w.r.t. corresponding goldstandard (GS) wordnets (French EWN and ManSloWNet). Results on the translation approach for Slovene are not shown, because they are not statistically significant (not enough data).

- the literal is not appropriate for that synset but is semantically very close to it, its hypernym or its hyponym (closely related); such cases can be considered as correct if more coarse-grained sense granularity is sufficient for a given application; for example, it might suffice to treat words, such as *ekipa* (*team*) and *skupina* (*group*) as synonyms in a particular HLT task;
- the literal is neither appropriate nor semantically related to the synset in question because it results from wrong sense disambiguation, wrong word alignment or wrong lemmatization (wrong).

The latter category contains real errors in the generated wordnets. Many of them (around 30% in Slovene data) are related to insufficient sense disambiguation at the stage of comparing wordnets in other languages. For example, the word *examination* can mean a medical check-up. In this case, the correct Slovene translation is *preiskava*. But when the same English word is used for a school exam, it should be translated as *preverjanje znanja*, not as *preiskava*. However, the latter was aligned twice with the English word *examination* and with the Czech word *zkouška*, whose meanings include *school exam*. This leads to a non-empty intersection of synset ids in the Sl-Cz-En source, which assigns the *school exam* synset to *preiskava*. Many errors are also the consequence of wrong word alignment of the corpus. This happened a lot in cases where the order of constituents in noun phrases in one language is substantially different from the order in another language. For example, the English compound *member state*$_{\text{head}}$ is always translated in the opposite order as *država*$_{\text{head}}$ *članica* in Slovene and *état*$_{\text{head}}$ *membre* in French, and is thus likely to be misaligned.

Source	Present in GS	Absent in GS but correct	**Source prec.**	Closely related	Wrong
WOLF					
Fr-Cz-En	61.7%	13.8%	**75.5%**	10.9%	13.6%
Fr-Cz-Bg-En	67.8%	12.4%	**80.1%**	9.2%	10.7%
Fr-Ro-En	64.7%	15.4%	**80.1%**	8.1%	11.8%
Fr-Ro-Cz-En	70.6%	13.3%	**84.0%**	8.4%	7.6%
Fr-Ro-Cz-Bg-En	75.5%	13.2%	**88.7%**	6.8%	4.5%
Wikipedia	94.0%	4.1%	**98.1%**	0.8%	1.1%
Fr Wiktionary	69.8%	12.2%	**82.0%**	10.7%	7.2%
En Wiktionary	88.5%	6.5%	**95.0%**	4.0%	1.1%
Wikispecies	*90.5%*	-	***90.5%***	-	*9.1%*
EUROVOC descr.	67.6%	*8.1%*	***75.7%***	*16.2%*	*8.1%*
SloWNet					
Sl-Cz-En	53.4%	7.6%	**61.0%**	4.7%	34.3%
Sl-Cz-Bg-En	59.3%	6.8%	**66.1%**	4.2%	29.7%
Sl-Ro-Cz-En	57.7%	7.5%	**65.2%**	3.8%	31.0%
Sl-Ro-Cz-Bg-En	61.0%	7.3%	**68.4%**	4.0%	27.6%

Table 3. Manual evaluation of WOLF and SloWNet and precision of BCS synsets according to the source used for generation. Figures in italics are to be considered carefully, given the low number of (literal, synset id) pairs.

The third source of errors are lemmatization problems, much more common in Slovene than French because the Slovene tagger was trained on a smaller corpus. If a strange lemma is guessed by the lemmatization algorithm for an unknown wordform, it will most likely be filtered out by the following stages in our synset generation procedure. However, if a word is assigned a wrong but legitimate lemma, it will be treated as a possible synonym for a certain concept by our algorithm and therefore appear in the wrong synset. For example, if the word form *vode* (singular genitive form of the lemma *water*) is wrongly lemmatized as *vod* (Eng. *platoon*), it will be placed in all the *water* synsets, which is a serious error that reduces the usability of the resource. In French, some expressions with plural canonical forms, such as *affaires* (*(one's) stuff*) got lemmatized into singular (*affaire*, Eng. *affair, deal, case*), which is inappropriate for that synset.

5 Conclusion

This paper has presented the two new lexico-semantic resources (wordnets) that were created automatically and are freely available for reuse and extension. The results obtained show that the approach taken is promising and should be exploited further as it yields a network of wide-coverage and quite reliable synsets[3] that can be used in many HLT applications. Some issues are still outstanding, however, such as the gaps in the hirerarchy and word sense errors.

[3] We plan to assign to them confidence levels according to source-by-source evaluation.

Manual revision of the work is required for better performance of the resources in a real life setting and is being carried out. Both wordnets could be further extended by mapping polysemous Wikipedia entries to PWN with a WSD approach similar to [15]. Next, lexicosyntactic patterns could be used to extract semantically related words from either the corpus [16] or Wikipedia [17]. Moreover, Wiktionaries start handling polysemy to some extent, including by differentiating translations according to senses defined by short gloses.

References

1. Vossen, P. (éd.): EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer, Dordrecht (1999)
2. Tufiş, D.: Balkanet design and development of a multilingual balkan wordnet. Romanian Journal of Information Science and Technology **7**(1–2) (2000)
3. Farreres, X., Rigau, G., Rodrguez, H.: Using WordNet for building WordNets. In: Proceedings of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, Canada (1998)
4. Barbu, E., Mititelu, V.B.: Automatic building of Wordnets. In: Proceedings of RANLP '05, Borovets, Bulgaria (2006)
5. Jacquin, C., Desmontils, E., , Monceaux, L.: Proc. of cicling'07 (lncs 4394). In: French EuroWordNet Lexical Database Improvements. (2007)
6. Pianta, E., Bentivogli, L., Girardi, C.: Multiwordnet: developing an aligned multilingual database. In: Proc. of the 1st Global WordNet Conf., Mysore, India (2002)
7. Resnik, P., Yarowsky, D.: A perspective on word sense disambiguation methods and their evaluation. In: ACL SIGLEX Workshop Tagging Text with Lexical Semantics: Why, What, and How?, Washington, D.C., United States (1997)
8. Ide, N., Erjavec, T., Tufiş, D.: Sense discrimination with parallel corpora. In: Proc. of ACL'02 Workshop on Word Sense Disambiguation. (2002)
9. Diab, M.: The feasibility of bootstrapping an arabic wordnet leveraging parallel corpora and an english wordnet. In: Proc. of the Arabic Language Technologies and Resources. (2004)
10. Fišer, D.: Leveraging parallel corpora and existing wordnets for automatic construction of the Slovene Wordnet. In: Proc. of L&TC'07, Poznań, Poland (2007)
11. Fišer, D., Sagot, B.: Proc. of Ontolex '08. In: Building a free French wordnet from multilingual resources. (2008) (to appear).
12. Ralf, S., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D.: The JRC Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proc. of LREC'06. (2006)
13. Tiedemann, J.: Combining clues for word alignment. In: Proc. of EACL'03, Budapest, Hungary (2003)
14. Erjavec, T., Fišer, D.: Building Slovene WordNet. In: Proc. of LREC'06, Genoa, Italy (2006)
15. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In: Proc. of Advances in Web Intelligence (LNAI 3528). (2005)
16. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proc. of COLING 1992, Nantes, France (1992)
17. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatic extraction of semantic relationships for wordnet by means of pattern learning from wikipedia. In: Proc. of NLDB 2005 (LNCS 3513), Alicante, Spain (2005)