WN-Toolkit:
Automatic generation of WordNets following the expand model

Antoni Oliver
Universitat Oberta de Catalunya
Barcelona - Catalonia - Spain
aoliverg@uoc.edu

Abstract

This paper presents a set of methodologies and algorithms to create WordNets following the expand model. We explore dictionary and BabelNet based strategies, as well as methodologies based on the use of parallel corpora. Evaluation results for six languages are presented: Catalan, Spanish, French, German, Italian and Portuguese. Along with the methodologies and evaluation we present an implementation of all the algorithms grouped in a set of programs or toolkit. These programs have been successfully used in the Know2 Project for the creation of Catalan and Spanish WordNet 3.0. The toolkit is published under the GNU-GPL license and can be freely downloaded from http://lpg.uoc.edu/wn-toolkit.

1 Introduction

WordNet (Fellbaum, 1998) is a lexical database that has become a standard resource in Natural Language Processing research and applications. The English WordNet (PWN - *Princeton WordNet*) is being updated regularly, so that its number of synsets increases with every new version. The current version of PWN is 3.1, but in our experiments we are using the 3.0 version because is the latest one available for download at the time of performing the experiments.

WordNet versions in other languages are also available. On the Global WordNet Association[1] website, a comprehensive list of WordNets available for different languages can be found. The Open Multilingual WordNet project (Bond and Kyonghee, 2012) provides free access to WordNets in several languages in a common format. We have used the WordNets from this project for Catalan (Gonzalez-Agirre et al., 2012) , Spanish (Gonzalez-Agirre et al., 2012) , French (WOLF) (Sagot and Fišer, 2008) , Italian (Multiwordnet) (Pianta et al., 2002) and Portuguese (OpenWN-PT) (de Paiva and Rademaker, 2012) . For German we have used the GermaNet 7.0 (Hamp and Feldweg, 1997), freely available for research. In Table 1, the sizes of all these WordNets are presented along with the size of the PWN.

	Synsets	Words
English	118.695	206.979
Catalan	45.826	46.531
Spanish	38.512	36.681
French	59.091	55.373
Italian	34.728	40.343
Portuguese	41.810	52.220
German	74.612	99.529

Table 1: Size of the WordNets

2 The expand model

According to (Vossen, 1998), we can distinguish two general methodologies for WordNet construction: (i) the *merge model*, where a new ontology is constructed for the target language; and (ii) the *expand model*, where variants associated with PWN synsets are translated using different strategies.

2.1 Dictionary-based strategies

The most commonly used strategy within the expand model is the use of bilingual dictionaries. The main difficulty faced is polysemy. If all the variants were monosemic, i.e., if they were assigned to a single synset, the problem would be simple, as we would only need to find one or more translations for the English variant. In Table 2 we can see the degree of polysemy in PWN 3.0. As we can see, 82.32% of the variants of the PWN are monosemic, as they are assigned to a single synset.

It is also worth observing the percentage of monosemic variants that are written with the first

[1] www.globalwordnet.org

N. synsets	variants	%
1	123.228	82.32
2	15.577	10.41
3	5.027	3.36
4	2.199	1.47
5+	3.659	2.44

Table 2: Degree of polysemy in PWN 3.0

letter in upper case (probably corresponding to proper names) and in lower case. In Table 3, we can see the figures.

	variants	%
upper case	84.714	68.75
lower case	38.514	31.25

Table 3: Number of monosemic variants with the first letter in uppercase or lowercase

These figures show us that a large percentage of a target WordNet can be implemented using this strategy. We must bear in mind, however, that using this methodology, we would probably not be able to obtain the most frequent variants, as common words are usually polysemic.

The Spanish WordNet (Atserias et al., 1997) in the EuroWordNet project and the Catalan WordNet (Benítez et al., 1998) were constructed using dictionaries.

With the dictionary-based strategy we will only be able to get target language variants for synsets having monosemic English variants, i.e. English words assigned to a single synset.

2.2 Babelnet

BabelNet (Navigli and Ponzetto, 2010) is a semantic network and ontology created by linking Wikipedia entries to WordNet synsets. These relations are multilingual through the interlingual relations in Wikipedia. For languages lacking the corresponding Wikipedia entry a statistical machine translation system is used to translate a set of English sentences containing the synset in the Semcor corpus and in sentences from Wikipedia containing a link to the English Wikipedia version. After that, the most frequent translation is detected and included as a variant for the synset in the given language.

Similarly to WordNet, BabelNet groups words in different languages into sets of synonyms, called Babel synsets. Babelnet also provides definitions or glosses collected from WordNet and Wikipedia. For cases where the sense is also available in WordNet, the WordNet synset is also provided. We can use Babelnet directly for the creation of WordNets for the languages included in Babelnet (English, Catalan, Spanish, Italian, German and French). For other languages, we can also exploit Babelnet through the Wikipedia's interlingual index.

Recently Babelnet 2.0 was released. This version includes 50 languages and uses information from the following sources: (i) Princeton WordNet, (ii) Open Multilingual WordNet, (iii) Wikipedia and (iv) OmegaWiki. a large collaborative multilingual dictionary.

Prelimiary results using this new version of Babelnet will be also shown in section 3.3.4.

With the Babelnet-based strategy we can get the target language variants for synsyets having both monosemic and polisemic English variants, that is, English words assigned to one or more synsets.

2.3 Parallel corpus based strategies

In some previous works we presented a methodology for the construction of WordNets based on the use of parallel bilingual corpora. These corpora need to be semantically tagged, the tags being PWN synsets, at least in the English part. As this kind of corpus is not easily available we explored two strategies for the automatic construction of these corpora: (i) by machine translation of sense-tagged corpora (Oliver and Climent, 2011), (Oliver and Climent, 2012a) and (ii) by automatic sense tagging of bilingual corpora (Oliver and Climent, 2012b).

Once we have created the parallel corpus, we need a word alignment algorithm in order to create the target WordNet. Fortunately, word alignment is a well-known task and several freely available algorithms are available. In previous works we have used Berkeley Aligner (Liang et al., 2006). In this paper we present the results using a very simple word alignment algorithm based on the most frequent translation. This algorithm is available in the WN-Toolkit.

With the parallel corpus based strategy we can get the target language variants for synsyets having both monosemic and polisemic English variants, that is, English words assigned to one or more synsets.

2.3.1 Machine translation of sense-tagged corpora

For the creation of the parallel corpus from a monolingual sense-tagged corpus, we use a machine translation system to get the target sentences. The machine translation system must be capable of performing a good lexical selection, that is, it should select the correct target words for the source English words. Other kinds of translation errors are less important for this strategy.

2.3.2 Automatic sense-tagging of parallel corpora

The second strategy for the creation of the corpora is to use a parallel corpus between English and the target language and perform an automatic sense tagging of the English sentences. Unfortunately word sense disambiguation is a highly error-prone task. The best WSD systems for English using WordNet synsets achieve a precision score of about 60-65% (Snyder and Palmer, 2004; Palmer et al., 2001). In our experiments we have explored two options: (i) the use of Freeling and UKB (Padró et al., 2010b) and (ii) Word Sense Disambiguation of multilingual corpora based on the sense information of all the languages (Shahid and Kazakov, 2010).

We have used *Freeling* (Padró et al., 2010a) and the integrated *UKB* module (Agirre and Soroa, 2009) to add sense tags to a fragment of the DGT-TM corpus (Steinberger et al., 2012). Before using this algorithm we have evaluated its the precision by means of automatically sense tag some sense tagged corpora: Semcor, Semeval2, Semeval3 and the Princeton WordNet Gloss Corpus (PWGC). After the automatic sense-tagging is performed, the tags are compared with those in the manually sense tagged-version. In Table 4 we can see the precision figure for each corpus and pos. As we can see, there is a great difference in precision. This difference can be explained by the complimentary values given in the table: the degree of ambiguity in the corpus and the percentage of open class words that are tagged in the corpus. As we can observe, the better precision value is achieved by the PWGC, having the smaller degree of ambiguity and the smaller percentage of tagged words. By contrast, the worse precision is achieved by the Semeval3 corpus, which has the highest degree of ambiguity and the highest percentage of tagged words.

We have also explored a word sense disambiguation strategy based on the sense information provided by a multilingual corpus, following the idea of (Ide et al., 2002). We have used the DGT-TM Corpus (Steinberger et al., 2012) in six languages: English, Spanish, French, German, Italian and Portuguese. We have sense tagged all the languages with no sense disambiguation, that is, giving all the possible senses to all the words in the corpus present in the WordNet versions for these languages. With all this sense information the Word Sense Disambiguation task consists of comparing the synsets in all languages for the same sentence, and taking the sense appearing the most times. Using this strategy some degree of ambiguity is still present after disambiguation. For example, for English the average number of synsets for tagged words before disambiguation is 5.96 (16.05% of the tagged words are unambiguous), and, after disambiguation, this figure is reduced to 2.46 (55.5% of the tagged words are unambiguous).

We have manually evaluated a small portion of this disambiguation strategy for the English DTG-TM corpus, obtaining a precision of 51.25%, very similar to the worst results for the Freeling+UKB strategy. One of the problems of the practical use of the multilingual word sense disambiguation strategy is the sensitivity of the methodology on the degree of development of the target WordNets. It is very important that the target WordNets used for tagging the target language corpora have registered all the senses for a given word. If this is not the case, we will get the wrong results.

3 The WN-Toolkit

3.1 Toolkit description

The toolkit we present in this paper collects several programs written in Python. All programs must be run in a command line and several parameters must be given. All programs have the option -h to get the required and optional parameters. The toolkit also provides some free language resources. The toolkit is divided in the following parts: (i) Dictionary-based strategies; (ii) Babelnet-based strategies, (iii) Parallel corpus based strategies and (iv) Resources, such as freely available lexical resources, pre-processed corpora, etc.

The *toolkit* can be freely downloaded from `http://lpg.uoc.edu/wn-toolkit`.

In the rest of this section, each of these parts of the toolkit are presented, along with the results of the experiments of WordNet extraction for the following languages: Catalan, Spanish, French, German, Italian and Portuguese. The evaluation of the

	Ambiguity	% tagged w.	Global	Nouns	Verbs	Adjectives	Adverbs
Semcor	7.61	84.24	51.99	58.64	40.68	61.57	68.91
Senseval 2	5.48	88.88	59.77	70.55	31.49	62.82	66.28
Senseval 3	7.84	89.44	51.82	57.08	42.46	59.72	100
PWGC	4.72	65.9	85.56	84.74	80.09	89.74	92.16

Table 4: Precision figures of the Freeling's implementation of UKB algorithm for four English Corpora

results is performed automatically using the existing versions of these WordNets. We compare the variants obtained for each synset in the target languages. If the existing version of WordNet for the given languages has the same variant for this synset, the result is evaluated as correct. If the existing WordNet does not have any variant for the synset, this result is not evaluated. This evaluation method has a major drawback: as the existing WordNets for the target languages are not complete (some variants for a given synset are not registered), some correct proposals can be evaluated as incorrect. For each strategy we have manually evaluated a subset of the variants evaluated as incorrect and those not evaluated for Catalan or Spanish. Crrected precision figures are presented for these languages.

3.2 Dictionary-based strategies

3.2.1 Introduction

Using this strategy we can obtain variants only for the synsets having monosemic English variants. We can translate the English variants using different kinds of dictionaries (general, encyclopedic and terminological dictionaries). We then assign the translations to the synset of the target language WordNet.

The WN-Toolkit provides several programs for the use of this strategy:

- **createmonosemicwordlist.py**: for the creation of the lists of monosemic words of the PWN. Alternatively, it is possible to use the monosemic word lists corresponding to the PWN version 3.0 distributed with the *toolkit*.

- **wndictionary.py**: using the monosemic word list of the PWN and a bilingual dictionary this program is able to create a list of synsets and the corresponding variants in the target language.

- **wiktionary2bildic.py**: this program creates a bilingual dictionary suitable for use with the program wndictionary.py from the xml dump files of Wiktionary[2].

- **wikipedia2bildic.py**: this program creates a bilingual dictionary suitable for the use with the program wndictionary.py from the xml dump files of the Wikipedia[3].

- **apertium2bildic.py**: this program creates a bilingual dictionary suitable for the use with the program wndictionary.py from the transfer dictionaries of the open source machine translation system Apertium[4] (Forcada et al., 2009). This resource is useful for Basque, Catalan, Esperanto, Galician, Haitian Creole, Icelandic, Macedonian, Spanish, Welsh and Icelandic, as there are available linguistic data for the translation system between English and these languages.

- **combinedictionary.py**: this program allows for the combination of several dictionaries, creating a dictionary with all the information from every dictionary, eliminating the repeated entries.

3.2.2 Experimental settings

We have used this strategy for the creation of WordNets for the following 6 languages: Catalan, Spanish, French, German, Italian and Portuguese. We have used Wiktionary and Wikipedia for all these languages and we have explored the use of additional resources for Catalan and Spanish. In Table 5 we can see the number of entries of the dictionaries created with the *toolkit* for all six languages using Wiktionary and Wikipedia.

	Wiktionary	Wikipedia
cat	9,979	31,578
spa	26,064	106,665
fre	30,708	142,142
deu	29,808	164,463
ita	20,542	77,736
por	15,280	42,653

Table 5: Size of the dictionaries

[2] www.wiktionary.org

[3] www.wikipedia.org

[4] http://apertium.org

3.2.3 Results and evaluation

In Table 6 we can see the results of the evaluation of the dictionary-based strategy using Wiktionary. The number of variants obtained depends on the Wiktionary size for each of the languages and ranges from 5,081 for Catalan to 18,092 for German. The automatic calculated precision ranges from 48.09% for German to 84.8% for French. This precision figure can be strongly influenced by the size of the reference WordNets, and more precisely on the number of variants for each synset. In the column *New variants* we can see the number of obtained variants for synsets not present in the target reference WordNet.

	Var.	**Precision**	**New var.**
cat	5,081	78.36	1,588
spa	14,990	50.93	8,570
fre	16,424	84.80	1,799
deu	18,092	48.09	12,405
ita	10,209	75.45	3,369
por	7,820	80.71	1,104

Table 6: Evaluation of the dictionary based strategy using Wiktionary

In Table 7 the results for the acquisition of WordNets from the Wikipedia as a dictionary are presented. The precision values are calculated automatically. The number of obtained variants is lower than the previous results from the Wikitionary.

	Var.	**Precision**	**New var.**
cat	290	63.29	132
spa	607	63.19	463
fre	654	71.49	177
deu	766	24.14	737
ita	361	52.17	292
por	315	72.93	85

Table 7: Evaluation of the dictionary based strategy using Wikipedia

We have extended the dictionary-based strategy for Catalan using the transfer dictionary of the open source machine translation system Apertium along with Wikipedia and Wiktionary. The resulting combined dictionary has 65,937 entries. This made it possible to create a new WordNet with 11,970 entries with an automatic calculated precision of 75.75%. We have manually revised 10% of the results for Catalan and calculated a corrected precision of 92.86% (most of the non-evaluated variants were correct and some of those evaluated as incorrect were correct too).

As we can see from Tables 6 and 7 the number of extracted variants from Wikipedia is smaller than the extracted from Wiktionary, although the dictionary extracted from Wikipedia is 3 or 4 times larger. This can be explained by the percent of encyclopedic-like variants in English WordNet, that can be calculated counting the number of noun variants starting by a upper-case letter. Roughly 30% of the nouns in WordNet are encyclopaedic variants, and this means about the 20% of the overall variants.

3.3 Babelnet-based strategies

3.3.1 Introduction

The program babel2wordnet.py allows us to create WordNets from the Babelnet glosses file. This program needs as parameters the two-letter code of the target language and the path to the Babelnet glosses file. With these two parameters, the program is able to create WordNets only for the languages present in Babelnet (in fact the program simply changes the format of the output). The program also accepts an English-target language dictionary created from Wikipedia (using the program wikipedia2bildic.py). This parameter is mandatory for target languages not present in Babelnet, and optional for languages included in Babelnet. The program also accepts as a parameter the $data.noun$ file of PWN, useful for performing caps normalization.

3.3.2 Experimental settings

For our experiments we have used the 1.1.1 version of Babelnet, along with the dictionaries extracted from Wikipedia as explained in section 3.2.2. We used the babel2wordnet.py program using the above-mentioned dictionary and the caps normalization option.

3.3.3 Results and evaluation

In Table 8 we can see the results obtained for Catalan, Spanish, French, German and Italian without the use of a complementary Wikipedia dictionary. Note that no values are presented for Portuguese, as this language is not included in Babelnet. For all languages, the precision values are calculated automatically taking the existing WordNets for these languages described in Table 1 as references.

Table 9 shows the results using the optional Wikipedia dictionary. Note that now results are presented for Portuguese, although this language

	Var.	Precision	New var.
cat	23,115	70.95	9,129
spa	31,351	76.80	19,107
fre	32,594	80.71	8,291
deu	32,972	52.10	27,243
ita	27,481	66.78	16.945
por	-	-	-

Table 8: Evaluation of the Babelnet-based strategy

is not present in Babelnet. These results are very similar with the results with no Wikipedia dictionary, except for Portuguese. This can be explained by the fact that Babelnet itself uses Wikipedia, so adding the same resource again (although a different version) leads to a very little improvements.

	Var.	Precision	New var.
cat	23,307	70.85	9,244
spa	31,604	76.61	19,301
fre	32,880	80.60	8,415
deu	33,455	51.79	27,651
ita	27,695	66.53	17,069
por	1,392	75.23	532

Table 9: Evaluation of the Babelnet-based strategy with Wikipedia dictionary

We have manually evaluated 1% of the results for Catalan and we obtained a corrected precision value of 89.17%

3.3.4 Preliminary results using Babelnet 2.0

In Table 10 preliminary results using the Babelnet 2.0 are shown. Please, note that precision values for Catalan, Spanish, French, Italian and Portuguese are marked with an asterisk, indicating that these values can not be considered as correct. The reason is simple, we are automatically evaluating the results with one of the resources used for constructing the Babelnet 2.0. Remember than one of the resoures for the construction of Babelnet 2.0 are the WordNet included in the Open Multilingual WordNet, the same WordNet used for automatic evaluation. Figures of new variants are comparable with the results obtained with the previous version of Babelnet.

	Var.	Precision	New var.
cat	84,519	*94.12	9,453
spa	81,160	*94.58	20,132
fre	34,746	*79,03	8,660
deu	35,905	49,43	29,522
ita	64,504	*93,83	17.782
por	28,670	*86.88	7,734

Table 10: Evaluation of the Babelnet-based strategy using Babelnet 2.0

Anyway, Babelnet 2.0 can be a good starting point for constructing WordNets for 50 languages. The algorithm for exploiting the Babelnet 2.0 for WordNet construction is also included in the WN-Toolkit. Please, note that this algorithm simply changes the format of the Babelnet file into the Open Multilingual Wordnet format.

3.4 Parallel corpus based strategies

3.4.1 Introduction

The WN-Toolkit implements a simple word alignment algorithm useful for the creation of WordNets from parallel corpora. The program, called synset-word-alignement.py, calculates the most frequent translation found in the corpus for each synset. We must bear in mind that the parallel corpus must be tagged with PWN synsets in the English part. The target corpus must be lemmatized and tagged with very simple tags (n for nouns; v for verbs; a for adjectives; r for adverbs and any other letter for other pos).

The synset-word-alignment program uses two parameters to tune its behaviour:

- The i parameter forces the first translation equivalent to have a frequency at least i times greater than the frequency of the second candidate. If this condition is not achieved, the translation candidate is rejected and the program fails to give a target variant for the given synset.
- The f parameter is the greater value for the ratio between the frequency of the translation candidate in the target part of the parallel corpus and the frequency of the synset in the source part of the parallel corpus.

3.4.2 Experimental settings

For our experiments we have used two strategies for the creation of the parallel corpus with sense tags in the English part.

- Machine translation of sense-tagged corpora. We have used two corpora: Semcor and Princeton WordNet Gloss Corpus. We have used Google Translate to machine translate these corpora to Catalan, Spanish, French, German, Italian and Portuguese.
- Automatic sense tagging of parallel corpora, using two WSD techniques: (i) WSD using multilingual information and (ii) Freeling + UKB. We have used a 118K sentences

fragment of the DGT-TM multilingual corpus (available in English, Spanish, French, German, Italian and Portuguese, but not in Catalan). We have chosen this number of sentences to have a corpus of a similar size to the Princeton WordNet Gloss Corpus

For our experiments we have set the parameter i to 2.5 and the parameter f to 5.

3.4.3 Results and evaluation

In Table 11 and 12 we can see the results for the use of machine translation of Semcor an PWGC. As we can see, the precision figures are very similar for both corpora, but the number of extracted variants is greater for the PWGC, due to the larger size of the corpus. We have manually evaluated 20% of the results for Catalan. In the case of Semcor we have calculated a corrected value of 94.74%, whereas for PWGC corpus we have obtained a corrected value of 96.18%.

	Var.	Precision	New var.
cat	2,001	87.63	449
spa	2,076	88.93	504
fre	1,844	91.83	142
deu	2,657	70.26	1,285
ita	858	93.81	66
por	2,064	84.14	324

Table 11: Evaluation of the parallel corpus based strategy: machine translation of Semcor corpus

	Var.	Precision	New var.
cat	4,744	87.87	1,125
spa	4,959	84.28	2,102
fre	4,598	91.63	510
deu	5,055	71.11	2,559
ita	4,870	88.68	904
por	4,845	86.26	871

Table 12: Evaluation of the parallel corpus based strategy: machine translation of PWGC corpus

In Table 13 and 14 we can see the results for the use of automatic sense tagging for the DGT-TM corpus using a multilingual strategy and Freeling+UKB. Here the precision figures are also similar for both strategies, but the number of extracted variants is greater for the Freeling+UKB strategy. The reason is that using Freeling and UKB we can disambiguate all the ambiguous words, while using the multilingual strategy we are not able to disambiguate all of them and in some cases some degree of ambiguity remains. For the extraction process we have only considered the fully disambiguated words.

	Var.	Precision	New var.
spa	313	75.35	171
fre	173	75.89	32
deu	207	36.54	155
ita	266	82.44	61
por	302	79.20	52

Table 13: Multilingual WSD of 118K sentences fragment of the DGT-TM corpus

	Var.	Precision	New var.
spa	1,155	79.71	386
fre	484	68.66	82
deu	609	24.72	431
ita	1,031	78.31	252
por	1,075	74.23	194

Table 14: Freeling + UKB of 118K sentences fragment of the DGT-TM corpus

In this case we have manually evaluated the results for Spanish as Catalan is not available in this corpus. For the multilingual strategy we have manually evaluated 100% of the results and calculated a corrected precision figure of 91.67%. For the Freeling + UKB results we have manually evaluated 25% of the results, obtaining a corrected precision of 88.94%.

If we analyse the results, we see that the extraction task has a much higher precision than the Word Sense Disambiguation strategies used to process the corpora. This may seem a little odd but we must bear in mind that we have used very restrictive values for the parameters i and f of the extraction program. These parameters allow us to extract only the best candidates, ensuring a good precision value for the extraction process, but a very poor recall value. It should be noted than for Spanish with the machine translation strategy we are getting 2,076 candidatesfor the Semcor Corpus and 4,959 for the Princeton Gloss Corpus, and we are now getting 313 candidates for the multilingual WSD strategy and 1,155 for the UKB WSD. If we force the extraction process to get 2,076 candidates, we obtain a precision value of 43.77% for the multilingual WSD strategy and 58.12% for UKB.

4 Resources

We are distributing some resources for several languages with the hope they can be useful to use the toolkit to create new WordNets or extend existing ones.

- Lexical resources: dictionaries created from Wiktionary, Wikipedia and Apertium transfer dictionaries.

- Preprocessed corpora: DGT-TM, Emea and United Nations Corpus from Opus[5] (Tiedemann, 2012). We have semantically-tagged the English part of the corpora with Freeling and UKB and lemmatized and tagged some of the target languages. We plan to preprocess other parallel corpora in the future.

5 Conclusions

We have presented the results of the automatic creation of WordNets for six languages using several techniques following the expand model. All these techniques are implemented in the freely available WN-Toolkit and have been successfully used for the expansion of the Catalan and Spanish WordNets under the Know2 project. The WordNets and the toolkit itself are being improved under the Skater Project. The successful use of this toolkit has also been reported for the Galician WordNet (Gómez Guinovart and Simões, 2013).

We can analyse the coincident extracted synsets and their associated precision for Catalan in Table 15. Here we have mixed the results for extended dictionary, Babelnet, translated PWGC and translated Semcor. The overall precision is 71.06% but, if we take into account the variants extracted using 2 or more methodologies, this precision rises up to 91.35%, although the number of extracted variants is drastically reduced.

Freq.	**Var.**	**Precision**	**New var.**
1+	35,142	71.06	13,997
2+	5,661	91.35	1,062
3+	1052	94.92	87
4+	135	96.06	8

Table 15: Evaluation of the repetition of the results for different strategies for Catalan

This combination of methodologies allows us to classify the extracted variants with an estimated precision value so we can obtain variants and give each variant a score. This score can be updated if the variant is obtained again using a different methodology or resource.

It's important to take into account the fact that the automatically-calculated precision value is very prone to errors, as, if a given synset having a variant lacks other possible variants and if those unregistered correct variants are extracted, the evaluation algorithm will consider them as incorrect. In Table 16 we can see the comparison between the automatic and corrected values of precision.

Strategy	**Lang.**	**% rev.**	$\mathbf{P}_{auto.}$	$\mathbf{P}_{corr.}$
Dictionaries	cat	10	75.75	92.86
Babelnet	cat	1	70.85	89.17
Semcor trad.	cat	20	87.63	94.75
PWGC trad.	cat	20	87.87	96.18
DGT-TM mult.	spa	100	75.35	91.67
DGT-TM UKB	spa	25	79.71	88.94

Table 16: Comparison of automatic and corrected precision figures

6 Future work

We plan to follow the development of the WN-Toolkit in the following directions: (i) change the script-oriented implementation of the current version to a class-oriented implementation allowing easy integration into another applications; (ii) increasing the number of integrated freely available resources and implementing a web query based use of some resources; (iii) developing a simple graphical user interface to facilitate its use and (iv) pre-processing and distributing more freely available corpora.

We also plan to use the toolkit to develop preliminary versions of WordNets for other languages.

Acknowledgments

This research has been carried out thanks to the Project *SKATER*, (TIN2012-38548-C06-01) of the Spanish Ministry of Science and Innovation

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41.

Jordi Atserias, Salvador Climent, Xavier Farreres, German Rigau, and Horacio Rodriguez. 1997. Combining multiple methods for the automatic construction of multi-lingual WordNets. In *Recent Advances in Natural Language Processing II. Selected papers from RANLP*, volume 97, pages 327–338.

Laura Benítez, Sergi Cervell, Gerard Escudero, Mònica López, German Rigau, and Mariona Taulé. 1998. Methods and Tools for Building the Catalan WordNet. In *In Proceedings of the ELRA Workshop on*

[5] http://opus.lingfil.uu.se/

Language Resources for European Minority Languages.

Francis Bond and Paik Kyonghee. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th International Global WordNet Conference, Matsue (Japan)*, pages 64–71.

Valeria de Paiva and Alexandre Rademaker. 2012. Revisiting a brazilian WordNet. In *Proceedings of the 6th Global Wordnet Conference*, Matsue (Japan).

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. The MIT press.

Mikel L. Forcada, Francis M. Tyers, and Gema Ramírez-Sánchez. 2009. The apertium machine translation platform: five years on. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 3–10.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0. In *Proceedings of the 6th Global WordNet Conference*.

Xavier Gómez Guinovart and Alberto Simões. 2013. Retreading dictionaries for the 21st century. In *Proceedings of the 2nd Symposium on Languages, Applications and Technologies (SLATE'13)*, pages 115–126.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic net for german. In *Proceedings of the ACL Workshor on Automatic Information Extraction and Building of Lexical and Sematic Resources for NLP Applications*, pages 9—15.

Nancy Ide, Tomaz Erjavec, and Dan Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8*, pages 61–66.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the HLT-NAACL '06*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 216–225, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1858704.

Antoni Oliver and Salvador Climent. 2011. Construcción de los wordnets 3.0 para castellano y catalán mediante traducción automática de corpus anotados semánticamente. In *Proceedings of the 27th Conference of the SEPLN, Huelva Spain*.

Antoni Oliver and Salvador Climent. 2012a. Building wordnets by machine translation of sense tagged corpora. In *Proceedings of the Global WordNet Conference, Matsue, Japan*.

Antoni Oliver and Salvador Climent. 2012b. Parallel corpora for wordnet construction. In *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (Cicling 2012). New Delhi (India)*.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010a. FreeLing 2.1: Five years of open-source language processing tools. In *LREC*, volume 10, pages 931–936.

Lluís Padró, Samuel Reese, Eneko Agirre, and Aitor Soroa. 2010b. Semantic services in freeling 2.1: Wordnet and UKB. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 21–24.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Developing an aligned multilingual database. In *Proc. 1st Int'l Conference on Global WordNet*.

Benoît Sagot and Darja Fišer. 2008. Building a free french wordnet from multilingual resources. In *Proceedings of OntoLex 2008*, Marrackech (Morocco).

Ahmad R. Shahid and Dimitar Kazakov. 2010. Retrieving lexical semantics from multilingual corpora. *Polibits*, 41:25–28.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 41–43, Barcelona (Spain).

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2012. Dgt-tm: A freely available translation memory in 22 languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 454–459.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, pages 2214–2218.

Piek Vossen. 1998. Introduction to Eurowordnet. *Computers and the Humanities*, 32(2):73–89.