Enlargement of the Croatian Wordnet using the WN-Toolkit

Scientific report

Antoni Oliver

1. Purpose of the visit

The purpose of the visit is the collaborative work for the enlargement of the Croatian Wordnet using the automatic construction techniques included in the WN-Toolkit. This toolkit is a set programs for the creation and enlargement of Wordnets using the expand model, that is, by translating the English variants of the Princeton Wordnet for English. The WN-Toolkit can be freely downloaded from http://sourceforge.net/projects/wn-toolkit/. The toolkit implements methodologies based on dictionaries, Babelnet and parallel corpora.

The Croatian WordNet has been developed under the Central and South-East European Resources (CESAR) project, funded by the European Commission (50%) and the University of Zagreb, Faculty of Humanities and Social Sciences (50%). The Croatian Wordnet has 10.031 synsets and 31.367 synset-variant pairs. The synset ID's are those of the Princeton WordNet for English v 3.0.

The Princeton WordNet for English version 3.0 has 117.659 synsets and 206.975 synset-variant pairs.

Data presented above show that the Croatian WordNet should be expanded in order to be a valuable resource for several Natural Language Processing tasks, for example Word Sense Disambiguation.

2. Description of the work carried out during the visit

During the visit we performed several experiments on Wordnet creation for Croatian. We used methodologies based on dictionaries, Babelnet and parallel corpora. In this section I will explain the methodologies and the resources and in the next section the evaluation results for each experiment will be presented along with the description of the evaluation methodology.

2.1. Dictionary based methodology

Description of the methodology

This strategy uses bilingual dictionaries to translate the English variants associated with each synset. This direct translation using dictionaries can be performed only on those English variants being monosemic, that is, variants associated to a single synset. About 82 % of the English variants in the Princeton WordNet 3.0 are monosemic. These figures shows us that a large percentage of a target Wordnet can be implemented using this strategy, but we would not be able to extract the most frequent variants, as common words are usually polisemic.

Resources

In the following table we can observe the dictionaries (English-Croatian) we have used for the experiments along with the number of entries.

Dictionary	**Website**	**Number of entries**
OmegaWiki	http://www.omegawiki.org/	1.692
Wiktionary	http://www.wiktionary.org/	29.216 / 7.437
Wikipedia	http://www.wikipedia.org/	70.387
Geonames	http://www.geonames.org/	1.353
Wikispecies	http://species.wikimedia.org/	1.785

The Wiktionary dictionary contains words in Croatian, Bosnian and Serbian, some of them written in Cyrillic. We have filtered the dictionary with the Croatian Morphological Dictionary in order to get a list of Croatian Words, so words in the Wiktionary dictionary not being in the Croatian Morphological dictionary are deleted from the dictionary. This way we have also deleted entries that are proper Croatian lemmas, but are not listed in the Croatian Morphological dictionary. Enlargement of Croatian Morphological dictionary is therefore necessary for obtaining even better results and more synset candidates.

Entries from the Wikipedia are all with the first letter in upper case. Once we have extracted the WordNet from Wikipedia we had to normalize the capitalization of the results. We have done this in an automatic way by comparing capitalization of entries from the Wikipedia with the capitalization of the variants of the same synset in the Princeton English WordNet.

Entries in the Wikispecies dictionary are with the first letter in upper case. In this case we have simple changed all to lower case.

2.2. Babelnet based strategies

BabelNet is a semantic network and a multilingual encyclopedic dictionary, with lexicographic and encyclopedic coverage of terms. Entries are connected in a very large network of semantic relations. BabelNet covers 50 languages, Croatian among them.

In this methodology we simply extract the data from the BabelNet file to get the target Wordnet.

For Croatian we have extracted a Wordnet with 12.949 synset-variant pairs. A caps normalization procedure has been done, as most of the entries in Babelnet 2 are capitalized.

2.3. Parallel corpus based methodologies

Description of the methodology

In order to extract Wordnets from a parallel corpus we need this parallel corpus to be semantically tagged with WordNet synsets in the English part. As these corpora are not easily available we use two strategies for the automatic construction of the required corpora:

- By machine translation of sense-tagged corpora. We use manually sense tagged English corpora (as Semcor, for example) and we automatically translate the English text into the target language. We are using Google Translate, as it is a statistical system capable to perform a quite good lexical selection task when translating, that is, in some cases is capable to select the correct translation of a polysemic word.
- By automatic sense-tagging of English-Croatian parallel corpora. To perform the sense-tagging we have used Freeling and UKB.

In both cases, we need to POS tag the Croatian text, getting both the lemma and the POS information. We have used Hunpos with a model for Croatian, and we have developed a program to get the associated lemma from the Croatian Morphological Lexicon.

Once we have these corpora, the task of extracting a WordNet can be seen as a word-alignment task. We have used GIZA++ to align the lemmatized parallel corpora and we have developed a script (that will be included in the WN-Toolkit) to extract the Wordnets from the aligned files.

Resources

In the following table we can see the information about the sense-tagged corpora for machine translation strategy.

Corpus	Website	Sentence pairs	English tokens	Croatian tokens
Semcor	http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor	37.176	794.748	721.282
PWGC	http://wordnet.princeton.edu/glosstag.shtml	113.404	1.529.105	1.303.386
Senseval 2	http://web.eecs.umich.edu/~mihalcea/downloads.html#sensevalsemcor	238	5.493	5.129
Senseval 3	http://web.eecs.umich.edu/~mihalcea/downloads.html#sensevalsemcor	300	5.530	5.022

And in the following table we can observe the information for the corpus used in the automatic sense-tagging strategy:

Corpus	Website	Sentence pairs	English tokens	Croatian tokens
Croatian English Parallel Corpus	http://metashare.elda.org/	62.566	1.790.041	1.590.637
EUBookshop	http://opus.lingfil.uu.se/EUbookshop.php	6.104	131.217	126.607
hrenWaC	http://nlp.ffzg.hr/resources/corpora/hrenwac/	47.475	1.282.007	1.152.552
SETIMES 2	http://opus.lingfil.uu.se/SETIMES2.php	205.910	4.629.877	4.662.863

3. Description of the main results obtained

3.1. Evaluation procedure

In order to automatically evaluate the results, we compare the obtained Wordnet with the existing Croatian WordNet. If we get some variant for a synset, we compare if in the Croatian Wordnet there is a variant for this synset, and if this variant is the same as the extracted one. If we got one of the variants in the reference Wordnet, the result is evaluated as correct. If there are some variants in the reference Wordnet, but not the one we extracted, this is evaluated as incorrect. If we don't have any variant in the reference Wordnet for the particular synset, the result remains unevaluated, that is, we don't take into account this obtained variant in the evaluation results. The automatic precision values obtained in this way tend to be lower than the real values. Sometimes we obtain a variant that is correct, but we have other correct variants for the same synset in the reference Wordnet. In these cases we evaluate our result as incorrect. On the other hand, as the reference Croatian Wordnet is not very big, we leave a lot

of obtained variants without evaluation. For this reason, for each experiment we have manually evaluated a subset of the non-evaluated and incorrect results in order to calculate a corrected value of precision.

We offer two values of corrected precision values:

- strict: we also have considered small errors (as capitalization, plural forms, etc.) as errors
- non-strict: we have considered small errors as correct.

3.2. Results for dictionary-based strategy

We have performed one extraction process using all the dictionaries at the same time. We have obtained the following results:

Total number synset-variant pairs	Automatic evaluated	New synset-variant pairs
7.247	1.156	6.091

And the following precision values:

Automatic precision value	Manually corrected strict precision	Manually corrected non-strict precision
70.33 %	84.49 %	90.72 %

3.3. Results for Babelnet-based strategy

We have obtained the following results:

Total number synset-variant pairs	Automatic evaluated	New synset-variant pairs
12.949	1.934	11.015

And the following precision values:

Automatic precision value	Manually corrected strict precision	Manually corrected non-strict precision
66.65 %	88.96 %	96.8 %

3.3. Results for parallel corpora based strategy: machine translation of manually sense-disambiguated corpora

We have obtained the following results using the following parameters:

- minimum frequency: 5
- minimum percent of the first candidate frequency vs. second candidate frequency: 50

Total number synset-variant pairs	Automatic evaluated	New synset-variant pairs
8.785	3.335	5.450

And the following precision values:

Automatic precision value	Manually corrected strict precision	Manually corrected non-strict precision
78.74 %	87.76 %	94.26 %

3.3. Results for parallel corpora based strategy: automatic sense-tagging of the English part of English-Croatian parallel corpora

We have obtained the following results using the following parameters:

- minimum frequency: 5
- minimum percent of the first candidate frequency vs. second candidate frequency: 50

Total number synset-variant pairs	Automatic evaluated	New synset-variant pairs
609	149	460

And the following precision values:

Automatic precision value	Manually corrected strict precision	Manually corrected non-strict precision
85.81 %	90.14 %	92.21 %

3.4. Main sources of errors

The manual revision of the results has allowed us to devise the main source for errors. We can highlight the following:

- For dictionary-based and Babelnet-based strategies one important source of errors is the capitalization of the entries. In some of the used dictionaries (for example Wikipedia and Wikispecies), all the entries begin with a capital letter, regardless they are proper or common names.
- For dictionary-based and Babelnet-based strategies other important source of errors are some entries in another forms other than nominative singular. Some of the dictionary entries are in nominative plural.
- For strategies based on parallel corpora (both machine translation of sense-tagged corpora and automatic sense-tagging of parallel corpora) the main source of errors are produced by the Croatian tagger. As stated earlier, we have used a simple Hunpos tagger with a model for Croatian and a simple script for adding the lemmata. This tagger is not able to cope with multiword expressions and is not able to attach the reflexive particle *se* of reflexive verbs to the lemma.
- For the strategy based on parallel corpora using machine translation, another important source of errors is the quality of the machine translation system. We have used Google Translate, a state-of-the-art machine translation system, so we don't expect to make any improvement in this aspect.
- For strategy based on parallel corpora using automatic word sense-disambiguation of the English part, one important source of errors is the word sense disambiguation, as it is a very difficult task. We have used a state-of-the-art word sense algorithm (Freeling+UKB), so we don't expect to make any improvement in this aspect.

4. Future collaboration with host institution

a. We plan to extend the experiments on automatic Wordnet extraction for Croatian using English-Croatian terminological resources, as IATE or Eurovoc, as well as some specialized dictionaries available on Internet.

b. For the use of the parallel corpus based strategy we need a POS tagger. For the experiments performed so far we have used Hunpos with a model for Croatian, and a script to add the lemma on the results using the Croatian Morphological Lexicon. We plan to create a better POS tagger by:

- Revising and improving the Croatian Morphological Lexicon
- Create a Freeling module for Croatian, using the improved Croatian Morphological Lexicon

and training Freeling with the Croatian National Corpus.

c. Once we have the Freeling for Croatian we plan to use the hrAcquis English-Croatian parallel corpus (648.238 segments).

5. Projected publications

We plan to present an abstract to the 29th International Conference Applied Linguistic Research and Methodology, that will be celebrated in Zadar (Croatia) in 24 – 26 April 2015. If the abstract is accepted, in the paper we will present the methodology and results of the experiments done during the research stay.

6. Other comments

The NetWords grant for short visits has been an excellent opportunity to start this work and to begin the collaboration between the Zagreb University and the Open University of Catalonia. As a result of this visit we plan to apply to a new NetWords grant for short visits for members of the Zagreb University to visit back the Open University of Catalonia and to continue this work.