Enlargement of the Croatian Wordnet using the WN-Toolkit

(and preliminary results for Slovene)

Antoni Oliver
(aoliverg@uoc.edu)
Universitat Oberta de Catalunya

Overview

- The WN-Toolkit
- The Expand Model
- Experiments and evaluation for Croatian
    - Dictionary-based strategy
    - Babelnet-based strategy
    - Parallel corpus based strategies
        - Machine Translation of sense-tagged corpora
        - Automatic sense tagging of parallel corpora
- Preliminary results for Slovene
- Conclusions
- Future work: 100WN Project

The WN-Toolkit

- A set of command-line programs in Python
- Some free language resources
- http://sourceforge.net/projects/wn-toolkit/
- GNU-GPL license

The Expand Model

- According to Vossen (1998) two methodologies for WordNet creation:
  - Merge Model: create a new ontology for the target language then link to PWN
  - Expand Model: translate the variants associated with PWN synset
- The WN-Toolkit uses Expand Model

Experiments and evaluation

- Experiments for Croatian
- Automatic evaluation using reference WordNets for target languages
- Partial manual evaluation
- Automatic evaluation tends to obtain lower results

The CroWN

- 10.031 synsets
- 31.367 synset-variant pairs
- XML-like format

```
<SYNSET><ID>ENG30-02400139-n</ID><POS>n</POS><SYNONYM><LITERAL sense="1">porodica
Bovidae</LITERAL><LITERAL sense="1">šupljorošci</LITERAL><LITERAL sense="1">porodica
šupljorožaca</LITERAL></SYNONYM><DEF>porodica unutar podreda …...goveda, bizoni, ovce i
koze</DEF><ILR type="hypernym">ENG30-01862557-n</ILR><ILR type="holo_member">ENG30-02398732-
n</ILR><STAMP></STAMP><BCS>3</BCS><SUMO>HoofedMammal<TYPE>+</TYPE></SUMO><DOMAIN>z
oology</DOMAIN></SYNSET>
```

- Transformed into OML Format

```
02400139-n     hrv:lemma       šupljorošci
```

The Princeton WordNet for English

- 117.659 synsets
- 206.975 synset-variant pairs

Automatic vs. manual evaluation

- No Ref. var. for the synset → Not evaluated
- Extr.Var. = Ref. Var. → Correct
- Extr.Var. ≠ Ref. Var. → Incorrect
  - It may be correct, but not present as a variant in the target WordNet. Example:
  - 15143477-n defunció mort death,dying,demise the time when something ends
  - It may be correct, but some error is present in the target WordNet. Example:
  - 14798039-n tetraclorur_de_carboni tertaclorur_de_carboni carbon_tetrachloride,carbon_tet,tetrachloromethane,perchloromethane a colorless nonflammable liquid used as a solvent for fats and oils; because of its toxicity its use as a cleaning fluid or fire extinguisher has declined

Files for evaluation

- Non-evaluated

13901321-n	cijev	pipe,tube	a hollow cylindrical shape
02533282-v	provjeriti	check	make an examination or...
*01902274-n	konj	horseback	the back of a horse
?09400037-n	protona	proton	a stable particle with positive...

- Incorrect

10476086-n zatvorenik uznik,zatočenik,zarobljenik,osoba koja je zarob ljena i drži se u zatočeništvu prisoner,captive a person who is confined; especially a prisoner of war

*11672400-n wildflower divlji cvijet,nekultivirani cvijet koji raste u divljini wildflower,wild_flower wild or uncultivated flowering plant

?07725376-n graška grašak,sjemenka graška (biljke) pea seed of a pea plant used for food

#02197781-v naći naći se, nalaziti se,... u kakvo stanje ili kakvu situaciju find perceive oneself to be in a certain condition or place

Dictionary-based strategy

- We get variants for synsets having monosemic English variants
  - PWN: 00098939-n self-service
  - DICT: self-service n samoposluživanje
  - WN-HRV: 00098939-n samoposluživanje
- We use several free dictionaries and encyclopedias (OmegaWiki, Wiktionary, Wikipedia, Geonames, Wikispecies)

Size of the dictionaries (eng-hrv)

OmegaWiki	1.692
Wiktionary	7.437
Wikipedia	70.387
Geonames	1.353
Wikispaces	1.785
TOTAL	79.984

Results for dictionary-based strategy

- Automatic Evaluation

Total	7.247
Evaluated	1.156
Precision	70.33 %
Precision N	70.49 %
Precision V	66.10 %
Precision A	83.33 %
Precision R	-

- Corrected (10%-NE – 20%-I)

Automatic	Precision	Precision *
70.33 %	**84.49 %**	90.72 %

Results for Babelnet based strategy

- Automatic Evaluation

Total	12.949
Evaluated	1.934
Precision	66.65 %
Precision N	66.65 %
Precision V	-
Precision A	-
Precision R	-

- Corrected (10%-NE – 20%-I)

Automatic	Precision	Precision *
66.65 %	**88.96 %**	96.8 %

Parallel corpus based strategies

We need a parallel corpus English-Croatian.:

- English: with semantic tags (WN synsets)
- Croatian: T.L lemmatised and POStagged

The WN creation is a word-alignment problem

Parallel corpus based strategies

They wanted to touch the mystery.
They 01825237-v to 02127358-v the 05685538-n .

Oni su htjeli dirati misterij .
Oni|oni|Pp3mpn--n-n-- su|biti|Vcr3p htjeli|htjeti|Vmp-pm dirati|dirati|Vmn misterij|misterij|Ncfsa .|.|Z
Oni biti htjeti dirati misterij

01825237-v	htjeti
02127358-v	dirati
05685538-n	misterij

Statistical alignment allowing word reordering and other phenomena

Parallel corpus based strategies

No such corpora available.
Two strategies for creation:

- Machine translation of sense-tagged corpora
    - Google Translate
    - Semcor, WordNet Gloss Corpus, Senseval 2, Senseval 3
- Automatic sense tagging of parallel corpora
    - Automatic English WSD with Freeling and UKB
    - Croatian-English parallel Corpus hr-en-p, HrenWac, Eubookshop, SETIMES 2

Size of the corpora: MT

	Sentence pairs	Tokens eng	Tokens hrv
Semcor	37.176	796.076	722.860
PWGC	113.404	1.530.250	1.304.426
Senseval2	230	5.500	5.132
Senseval3	300	5.557	5.022

Machine translation of sense-tagged English Corpora

- Automatic Evaluation

Total	8.785
Evaluated	3.335
Precision	78.74 %
Precision N	79.10 %
Precision V	75.21 %
Precision A	89.93 %
Precision R	-

- Corrected (10%-NE – 20%-I)

Automatic	Precision	Precision *
78.74 %	**87.76 %**	94.26 %

Size of the corpora for A. WSD

	Sentence pairs	Tokens eng	Tokens hrv
hre-en_p.c	62.566	1.790.041	1.590.637
EUbookshop	6.104	131.217	126.607
hrenWac	47.475	1.282.007	1.152.552
SETIMES2	205.910	4.629.877	4.662.863

Automatic WSD of English-Croatian parallel corpora

- Automatic Evaluation

Total	609
Evaluated	149
Precision	85.91 %
Precision N	84.13 %
Precision V	87.71 %
Precision A	100 %
Precision R	-

- Corrected (100%-NE – 100%-I)

Automatic	Precision	Precision *
85.81 %	**90.14 %**	92.21 %

Overall results

- Automatic Evaluation

Total	23.567
Evaluated	5.315
Precision	71.03 %
Precision N	69.83 %
Precision V	74.19 %
Precision A	89.74 %
Precision R	-

- 18.252 new synset-variant pais
- Expected precision up to 85 %

Main source of errors

- Dictionary and Babelnet2
  - capitalization of the entries: In Wikipedia and Wikispecies. Improve automatic normalization
  - Some entries in another forms other than nominative singular (i.e. nominative plural)
- Parallel corpora
  - Tagger: simple one. (i.e.. reflexive verbs with *se*)
  - Machine translation
  - WSD

Preliminary results for Slovene

- SloWNet
  - 42.586 synsets
  - 74.130 synset-variant pairs

Results for dictionary-based strategy

- Automatic Evaluation

Total	7.370
Evaluated	5.421
Precision	66.09 %
Precision N	67.09 %
Precision V	60.53 %
Precision A	40.32 %
Precision R	24.32

Results for Babelnet 2

- Automatic Evaluation

Total	10.670
Evaluated	7.078
Precision	73.52 %
Precision N	73.52 %
Precision V	-
Precision A	-
Precision R	-

Overall

- Automatic Evaluation

Total	13.933
Evaluated	9.480
Precision	67.27 %
Precision N	67.86 %
Precision V	60.53 %
Precision A	40.32 %
Precision R	24.32

- New synset-variant pairs: 4.453

Conclusions

- Effective way to create or enlarge WN
- 100WN-Project

Future work

- Improve the WN-Toolkit
- Train Freeling for Croatian
  - Improving HML (GPL)
- HrAcquis parallel corpus (Croatian)
- Measure of completeness of a WordNet
- Manually revise all the results:
  - In a variant frequency order
- 100WN: collaborate in the creation or expansion of WordNets for other languages

Thank you very much

aoliverg@uoc.edu