Extending the WN-Toolkit: dealing with polysemous words in the dictionary-based strategy

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

Overview

- WN-Toolkit
- Improving dictionary-based strategy
- Omegawiki
- Wiktionary
- Database design
- Algorithm
- Evaluation procedure
- Results
- Learning the best combination
- Easy and quick revision of the results
- Future work
- Conclusions

WNToolkit

- A set of Python programs for WordNet automatic wordnet creation:
  - Dictionary-based strategy
  - Babelnet
  - Parallel corpus based strategy
    - Machine translation of sense-tagged corpora
    - Automatic sense tagging of parallel corpora
  - Algorithms for automatic evaluation of the results
- Some freely available language resources are distributed
- Available at: http://sourceforge.net/projects/wn-toolkit/

Improving the dictionary-based strategy

- In previous versions only monosemous English variants are translated using dictionaries
- Most of the English variants in WordNet are monosemous. But frequent words tend to be polysemous

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

Monosemous variants

- acid rain - 14517629-n
  - cat: pluja àcida
  - spa: lluvia ácida
  - cnm: 酸雨
  - ell: όξινη βροχή
  - nld: zure regen
  - ….

Polysemous variants

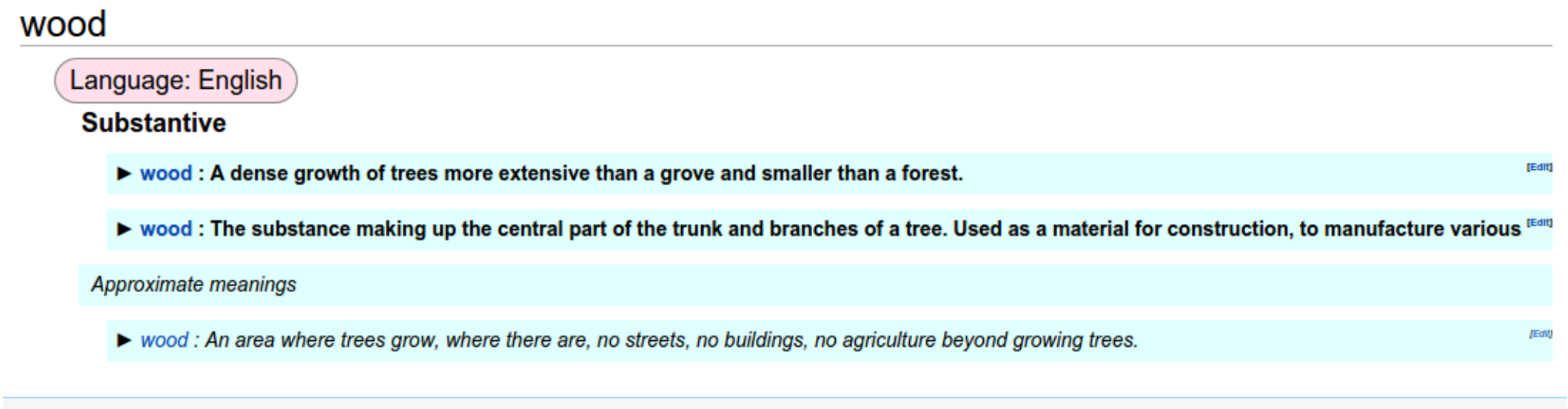
- wood: 15098161-n, 08438533-n
  - cat: fusta, bosc
  - spa: madera, bosque
  - cnm: 木 , 林
  - ell: ξύλο, δάσος
  - nld: hout, bos
  - ...

Improving the dictionary-based strategy

- In this stay I tried to implement algorithms for expanding this strategy to polysemous English words:
  - Using definitions both in WordNet and in the dictionaries
  - Using semantic relations both in WordNet and in the dictionaries
- I used OmegaWiki and Wiktionary
- For 25 languages:
  - 24 of them present in the OMW
  - Dutch

Omegawiki
http://www.omegawiki.org/

- http://www.omegawiki.org/Expression:wood



SQL Dump

- Omegawiki can be downloaded as a MySQL dump.
- The dump is loaded into a MySQL database
- All the information we need is selected and inserted into another database

- Omegawiki definitions and translations to cat:
  - A dense growth of trees more extensive than a grove and smaller than a forest. - bosc
  - The substance making up the central part of the trunk and branches of a tree. Used as a material for construction, to manufacture various items, etc. or as fuel. – fusta
- WordNet definitions and synsets
  - the hard fibrous lignified substance under the bark of trees - 15098161-n
  - the trees and other plants in a large densely wooded area - 08438533-n

- http://www.omegawiki.org/Expression:bank
- Omegawiki definitions and translations to spa and relations:
  - The sloping side of any hollow in the ground, especially when bordering a river. - margen, ribera
  - A financial institution where one can borrow money (upon which interest is due) or deposit money (in order to collect interest).- banco
- WordNet definitions and synsets
  - sloping land (especially the slope beside a body of water): 09213565-n
  - a financial institution that accepts deposits and channels the money into lending activities - 08420278-n

Relations

- Some entries in Omegawiki provide semantic relations
- http://www.omegawiki.org/Expression:citrus_fruit

- Comparing these relations with relations in WordNet we can disambiguate some entries
- Drawback: only few entries have relations:
  - 19,512 entries in OmegaWiki (41,174)
  - 11,083 entries in Wiktionary (out of 50,131)

Used relations

- In the experiments we used the following relations with omegawiki

hyponym	864
hypernym	2,896
holonym	121
meronym	26
antonym	182
TOTAL	3,448

Wiktionary

- We use the definitions in Wiktionary in a similar way as Omegawiki
- https://en.wiktionary.org/wiki/wood
- Wiktionary is distributed as a XML dump

```
 <title>wood</title>

   <ns>0</ns>

   <id>4648</id>

   …

   ====Noun====

   {{en-noun|~}}

   # {{context|uncountable|lang=en}} The substance making up the central part of the
   [[trunk]] and [[branches]] of a [[tree]]. Used as a material for construction, to
   manufacture various items, etc. or as fuel.

   #: {{usex|lang=en|This table is made of '''wood'''.&amp;amp;nbsp;&amp;amp;nbsp; There was lots of
   '''wood''' on the beach.}}

   …

   =====Translations=====

-  {{trans-top|substance}}

   * Abkhaz: {{t-needed|ab}}

   * Afrikaans: {{t+|af|hout}}

   * Alabama: {{t|akz|itto}}

   * Albanian: {{t|sq|dru}}

   * Amharic: {{t|am|እንጨት|tr=ʾənəč'ätə|sc=Ethi}}

   * Arabic: {{t+|ar|خشب|m|tr=xášab|sc=Arab}}
```

Relations in Wiktionary

- We are not using relations with wiktionary as we extracted a different set of relations, namely:

Derived term:	7,656
Related term:	5,040
TOTAL:	11,083

WNTK Dictionary Database

- The information in OmegaWiki and Wiktionary is stored in a MySQL database
- Tables:
  - definition
  - entry
  - relations
  - tagged_definition
  - target_definition
  - translations

Algorithm

- Select all translations along with the entry_id that allows to get:
  - English word
  - POS
  - Tagged definition
  - Relations
- Select all the synsets in WordNet for the English word and POS, and select:
  - Tagged definition
  - Relations

- For each synset:
  - We compare the tagged definitions and count the number of equal lemmas
  - We compare the relations and count the number of equal lemmas
  - We multiply each factor by a weight
- The synset with a higher value is attached to the target word

Evaluation procedure

- As 24 of the WordNets are in the OMW we can perform an automatic evaluation
- Having pairs of synset-variants (SV)
- If SV_extracted in OMW: CORRECT
- If SV_extracted not in OMW and Synset in OMW: INCORRECT
- If SV_extracted not in OMW and Synset not in OMW: NON EVALUATED

- If SV_extracted not in OMW and Synset in OMW: we evaluate as INCORRECT but in fact it can be CORRECT, being a new variant for this synset
- The automatic evaluation results tend to be lower than real values

Results OmegaWiki

lang	P (all, nodes)	All, nodes	P (all, des)	All, des	P (mono)	Mono	P (poly, nodes)	Poly, nodes	P (poly, des 1-5)	Poly, des 1-5
als/sqi	39.78	1,575	58.59	345	58.33	135	38.46	1,441	58.67	211
arb	33.23	10,747	49.35	2,403	49.14	1,237	31.24	9,511	49.52	1,167
bul	36.66	29,183	66.84	3,461	63.09	1,862	35.66	27,322	68.46	1,600
cat	51.72	9,582	70.76	2,859	69.81	1,672	47.02	7,911	72.06	1,188
cmn-Hant	11.09	13,165	27.39	2,293	30.45	1,051	9.59	12,115	25.44	1,243
dan	46.83	12,977	66.32	4,302	58.30	2,522	45.19	10,456	70.71	1,781
ell	33.60	13,821	51.80	3,766	52.34	2,009	30.21	11,813	51.29	1,758
eus	47.58	8,370	66.28	2,860	64.99	1,708	43.46	6,663	67.75	1,153
fin	33.81	23,089	62.17	5,924	64.25	3,343	28.66	19,747	59.47	2,582
fra	45.99	65,640	59.26	14,114	57.55	7,852	44.45	57,789	61.35	6,263
glg	64.01	3785	81.34	1269	83.6	752	49.78	3034	75.94	518
hrv*	51.59	4180	74.2	1314	78.89	701	43.18	3480	68.27	614
hrv	56.23	4180	79.9	1314	82.7	701	48.08	3480	76.37	614

- Some knowledge of the target language is needed in order to interpret the results:
  - hrv: in OmegaWiki the include “accents” to denote intonation that are not present in the regular written text (nor in the OMW reference wordnet): bȍlēst, Európa, bŕdo
  - spa: in OmegaWiki forms other than the lemma are included, but not in the reference OMW wordnet:
    - 14584110-n actínidos / 14584110-n actínido
    - 09605289-n adulto / 09605289-n adulta
  - Diacritics in arab….
  - ...

Learning the best combination

- The parameters are stored in a file so we can apply method for finding the best combination:

pluja àcida MONO 14517629-n

àcid POLY 14607521-n/2:1:0:0:0:0;02675657-n/0:0:0:0:0:0

actini MONO 14627655-n

acte POLY 06532095-n/1:0:0:0:0:0;00030358-n/1:0:0:0:0:0;07009640-n/0:0:0:0:0:0;06892016-n/0:0:0:0:0:0;07014029-n/0:0:0:0:0:0

agricultura POLY 01104406-n/0:0:0:0:0:0;00916464-n/0:0:0:0:0:0;08128964-n/0:0:0:0:0:0;08075287-n/0:0:0:0:0:0

agricultura POLY 01104406-n/0:0:0:0:0:0;00916464-n/0:0:0:0:0:0;08128964-n/0:0:0:0:0:0;08075287-n/0:0:0:0:0:0

aire POLY 14841267-n/2:0:1:0:0:0;08653314-n/0:0:0:0:0:0;04727214-n/0:0:0:0:0:0;11431754-n/0:0:0:0:0:0;08499057-n/0:0:0:0:0:0;14842703-n/0:0:0:0:0:0;07028373-n/1:0:0:0:0:0;06255354-n/0:0:0:0:0:0;00300441-n/0:0:0:0:0:0

Experiments for Catalan

Def.	Rel.	P des	Des	P noamb	Noamb	P amb	Amb
0	1	70.37	1794	69.81	1672	76.99	123
1	0	70.18	3041	69.81	1672	70.61	1116
1	2	70.19	3051	69.81	1672	70.63	1380
1	5	70.23	3051	69.81	1672	70.72	1380
2	1	70.15	3053	69.81	1672	70.54	1382
5	1	70.15	3054	69.81	1672	70.54	1383

Easy and quick revision of the results

- A couple of files can be provided to the revisors for easy and quick revision
- Nonevaluated:

14796969-n diòxid de carboni carbon_dioxide,CO2,carbonic_acid_gas a heavy odorless colorless gas formed during respiration and by the decomposition of organic substances; absorbed from the air by plants in photosynthesis

01262441-n desforestació deforestation,disforestation the removal of trees

13462989-n desertització desertification the gradual transformation of habitable land into desert; is usually caused by climate change or by destructive use of the land

08173515-n unió europea European_Union,EU,European_Community,EC,European_Economic_Community,EEC, Common_Market,European international organization of European countries formed after World War II to reduce trade barriers and increase cooperation among its members

....

Easy and quick revision of the results

- Incorrect:

00454237-n pesca pesca de canya angling fishing with a hook and line (and usually a pole)

00582388-n negoci ocupació occupation,business,job,line_of_work,line the principal activity in your life that you do to earn money

02233338-n escarabat cuca panera,escarabat de cuina cockroach,roach any of numerous chiefly nocturnal insects; some are domestic pests

11512818-n conductivitat elèctrica conducció,conductibilitat,conductivitat conduction,conductivity the transmission of heat or electricity or sound

14840092-n pols pols (partícules) dust free microscopic particles of solid material

01935395-n llambric cuc de terra earthworm,angleworm,fishworm,fishing_worm,wiggler,nightwalker,nightcrawler,crawler,dew_worm,red_worm terrestrial worm that burrows into and helps aerate soil; often surfaces when the ground is cool or wet; used as bait by anglers

07775375-n pescar peix fish the flesh of fish used as food

11482706-n boira boirina,broma,calitja mist a thin fog with condensation near the ground

11482706-n boirim boirina,broma,calitja mist a thin fog with condensation near the ground

….

Future work

- Finish the results for OmegaWiki (inminent)
- Run the experiments for Wiktionary
- Using Wikipedia
- Other dictionaries: Apertium transfer dictionaries, Wikispecies, some language specific and proprietary (under agreement)…
- Run the experiments for all the languages in the resources
- Agreement with universities/institutions to revise the results

Future work

- Optimize the database to faster performance
- Improve the parsers for Wiktionary, Wikipedia, Wikispecies….
- Extend the definition-based strategy using related words
- Compare and share the results with the Extended Open Multilingual Wordnet (Francis Bond)
- Improve and pack all the algorithms and resources in the new version of the WN-Toolkit.

Conclusions

- A simple and effective method for construction of wordnets from freely available dictionaries has been presented
- It can extract translations from English polysemous words using definitions and relations from WordNet and the dictionaries