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
  - Monosemous variants
  - Polysemous variants
- The Open Multilingual Wordnet
- Omegawiki
- Wiktionary
- WNTK dictionary database
- Algorithm
  - Disambiguation procedure
  - Weights
- Automatic evaluation procedure
- Results
  - Results for Omegawiki
  - Results for Wiktionary
- Comments on the results
- Optimization of the weights
- Easy and quick revision of the results
- Conclusions
- Future work

WNToolkit

- A set of Python programs for automatic wordnet creation following the expand model (translation of the variants)
- Several strategies
  - 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 were 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

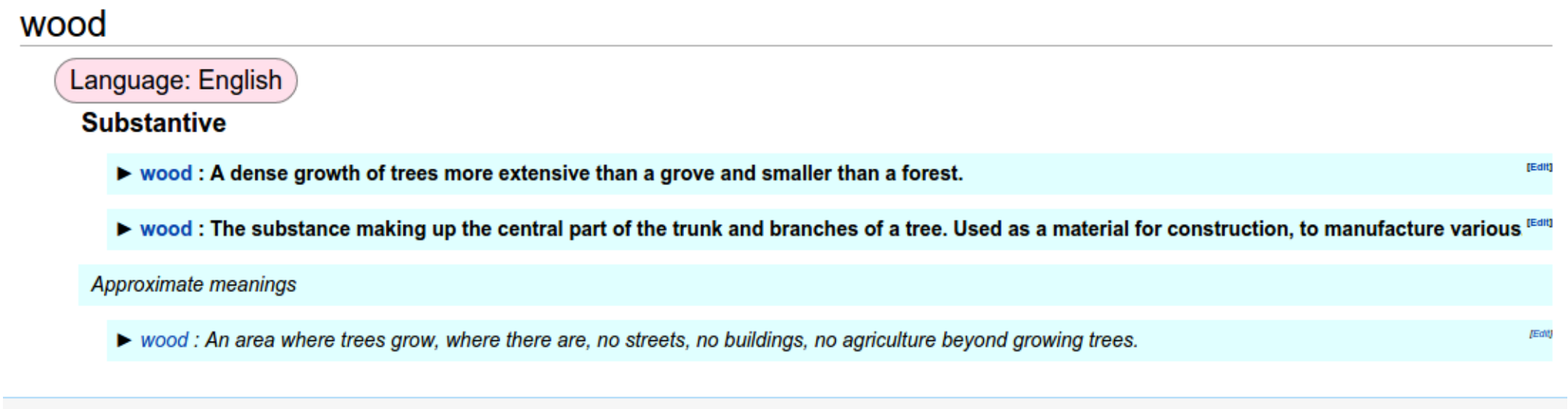
- Using definitions both in WordNet and in the dictionaries
- Using semantic relations both in WordNet and in the dictionaries
- OmegaWiki and Wiktionary
- For 24 languages in the OMW

The Open Multilingual Wordnet

Language	Code	Synsets	Words	Senses	Core
Albanian	sqi	4,676	5,990	9,602	31%
Arabic	arb	10,165	14,595	21,751	48%
Basque	eus	29,413	26,240	48,934	71%
Bulgarian	bul	4,999	6,783	9,056	100%
Catalan	cat	45,826	46,531	70,622	81%
Chinese	cmn	42,312	61,533	79,809	100%
Croatian	hrv	23,122	29,010	47,906	100%
Danish	dan	4,476	4,468	5,859	81%
Finnish	fin	116,763	129,839	189,227	100%
French	fra	59,091	55,373	102,671	92%
Galician	glg	19,312	23,124	27,138	36%
Greek	ell	18,049	18,227	24,106	57%
Hebrew	heb	5,448	5,325	6,872	27%
Indonesian	ind	38,085	36,954	106,688	94%
Italian	ita	35,001	41,855	63,133	83%
Japanese	jpn	57,184	91,964	158,069	95%
Norwegian N.	nno	3,671	3,387	4,762	66%
Norwegian B.	nob	4,455	4,186	5,586	81%
Polish	pol	36,054	61,393	88,889	66%
Portuguese	por	43,895	54,071	74,012	84%
Slovene	slv	42,583	40,233	70,947	86%
Spanish	spa	38,512	36,681	57,764	76%
Swedish	swe	6,796	5,824	6,904	99%
Thai	tha	73,350	82,504	95,517	81%

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

Relations in Omegawiki

- They use an open set of relations
- Relations appearing more than 50 times

relation	freq.
is part of theme	16,158
parent	11,980
child	11,776
broader terms	7,299
narrower terms	5,639
is spoken in	4,692
related terms	3,717
borders on	797
is written in	633
antonym	328
official language	226
capital	209
country	192
wordt gevolgd door	178
currency	165
holonym	183
demonym	122
flows through	110
dialectal variant	78
meronym	73
flows into	68
is practiced by a	61

Conversion of relations

Code OW	Relation OW	Relation S.
4	broader terms	hypernym
5	narrower terms	hyponyms
7574	antonym	antonym
375074	meronym	meronym
375078	holonym	holonym
-	translation into same language	synonym

- 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
- Comparing these relations with relations in WordNet we can disambiguate some entries

Wiktionary

- We use the definitions in Wiktionary in a similar way as Omegawiki
- Wiktionary is distributed as a XML dump
- It is not simple to parse it
- We are using Dbnary

Relations in Wiktionary

- The following relations are present in Wiktionary:
  - Synonym
  - Hypernym
  - Hyponym
  - Meronym
  - Holonym

WNTK Dictionary Database

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

Algorithm

- Select all entry ids, English words, target language words and pos from the table entry
  - For the Eng word and POS we search PWN for all synsets
  - It it belongs to only one synset MONOSEMOUS the target words are attached
  - If it belongs to several synsets POLYSEMOUS → DISAMBIGUATION PROCEDURE

Disambiguation procedure

- Select all relations from dictionaries
- Select all relations from PWN
- For each synset we count the coincident related words. A specific weight is given.
- Select the tagged definitions from dictionaries
- Select the the tagged definition from PWN
- For each synset we count coincident open class lemmata. A weight is applied.
- The synset with the higher score is attached to the target language word

Weights

- A set of weights has to be defined
  - For each relation
  - For coincident open class lemmata
- In our experiments:
  - 5 for relations
  - 1 for coincident open class lemmata
- A procedure for optimization of weights is presented

Automatic 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

Lang.	Omegawiki			Wiktionary		
	C	I	N	C	I	N
sqi	58	45	249	296	207	2,263
arb	68	902	1,507	289	2,561	6,128
eus	1,339	694	881	1,192	532	635
bul	516	256	2,688	866	1,680	10,514
cat	1,671	680	554	5,697	2,915	3,881
cmn	857	526	1,344	3,640	8,140	14,775
hrv	785	274	287	2,151	7,120	4,757
dan	535	269	3,612	964	757	774
fin	3,778	2,309	18	17,551	21,325	127
fra	7,440	5,168	1,963	16,545	9,713	5,110
glg	589	134	561	1,579	498	2,328
ell	1,041	948	1,852	2,697	2,863	9,606
heb	29	575	2,018	133	1,390	5142
ind	919	484	259	1,704	1,383	758
ita	5,627	3,814	4,471	8,671	6,375	7,836
jpn	2,871	1,306	650	9,786	8,374	3,792
nno	70	17	517	326	222	2,668
nob	480	242	3,063	394	277	2,844
pol	2,348	1,310	1,434	6,133	4,402	5,817
por	4,832	1,810	474	12,892	7,741	5,410
slv	1,663	888	445	2,566	1,790	638
spa	4,088	4,567	8,525	6,179	7,155	15,274
swe	1,104	699	4,640	2,238	2,437	16,007
tha	733	464	85	1,639	1,632	330

Results for Omegawiki

Lang.	All no dis.		All dis.		Non ambiguous		Amb. no dis.		Amb. dis.	
	Entries	Precision	Entries	Precision	Entries	Precision	Entries	Precision	Entries	Precision
sqi	1,466	40.18	353	56.31	135	58.33	1,332	38.76	219	55.7
arb	9,191	4.19	2,478	7.01	1,237	9.83	7,955	3.34	1,242	4.97
eus	7,934	48.03	2,915	65.86	1,708	64.99	6,227	43.77	1,208	66.8
bul	29,183	36.66	3,461	66.84	1,862	63.09	27,322	35.66	1,600	68.46
cat	8,531	53.57	2,906	71.08	1,673	69.76	6,859	48.74	1,234	72.81
cmn-Hans	11,924	26.88	2,728	61.97	1,269	68.88	10,656	22.39	1,460	57.71
hrv	4,180	51.59	1,347	74.13	701	78.89	3,480	43.18	647	68.4
dan	11,935	48.38	4,417	66.54	2,523	58.3	9,413	46.8	1,895	70.73
fin	20,134	36.02	6,106	62.07	3,342	64.24	16,793	30.4	2,765	59.44
fra	53,499	48.3	14,572	59.01	7,850	57.48	45,650	46.75	6,723	60.74
glg	3,483	65.34	1,285	81.47	753	83.43	2,731	51.16	533	76.85
ell	12,838	34.61	3,842	52.34	2,009	52.29	10,830	31.09	1,834	52.38
heb	9,199	2.56	2,623	4.8	1,347	4.37	7,853	2.22	1,277	5.11
ind	5,589	48.06	1,663	65.5	852	64.68	4,738	44.86	812	66.33
ita	85,324	32.05	13,913	59.6	6,614	59.82	78,711	29.41	7,300	59.41
jpn	14,994	40.48	4,828	68.73	2,694	71.59	12,301	33.16	2,135	65.32
nno	1,379	59.89	605	80.46	376	80.0	1,004	55.92	230	80.7
nob	10,555	47.0	3,786	66.48	2,196	58.61	8,360	45.21	1,591	70.5
pol	16,417	41.99	5,093	64.19	2,876	64.67	13,542	35.37	2,218	63.55
por	26,301	52.52	7,117	72.75	3,761	69.16	22,541	48.36	3,357	77.10
slv	9,136	49.68	2,997	65.19	1,607	61.59	7,530	47.1	1,391	69.21
spa	68,884	31.65	17,181	47.23	8,874	41.86	60,011	30.55	8,308	51.41
swe	21,626	40.05	6,444	61.23	3,535	63.17	18,092	35.67	2,910	59.81
tha	4,065	33.08	1,283	61.24	677	59.87	3,389	27.12	607	62.72

Results for Wiktionay

Lang.	All no dis.		All dis.		Non ambiguous		Amb. no dis.		Amb. dis.	
	Entries	Precision	Entries	Precision	Entries	Precision	Entries	Precision	Entries	Precision
sqi	11,510	43.02	2,767	58.85	1,251	59.03	10,260	41.91	1,517	58.77
arb	37,540	6.75	8,980	10.14	4,431	12.3	33,110	6.21	4,550	8.79
eus	9,359	50.7	2,360	69.14	1,244	69.89	8,116	47.41	1,117	68.43
bul	59,664	25.23	13,061	34.01	5,690	34.95	53,975	24.53	7,372	33.63
cat	53,737	52.1	12,494	66.15	6,597	68.86	47,141	49.5	5,898	63.22
cmn	102,130	18.98	26,519	31.47	30	36.36	88,589	16.79	14	25.0
hrv	62,765	17.4	14,029	23.2	6,399	25.57	56,367	16.17	7,631	20.99
dan	43,052	39.95	9,469	56.01	4,866	62.01	38,187	38.4	4,604	53.65
fin	174,743	26.22	39,004	45.15	19,958	54.95	154,786	22.51	19,047	34.88
fra	119,160	53.91	31,369	63.01	17,802	66.24	101,359	51.67	13,568	58.51
glg	17,745	59.92	4,406	76.02	2,261	77.95	15,485	52.99	2,146	72.66
ell	67,014	32.3	15,168	48.51	7,408	55.4	59,607	29.85	7,761	43.35
heb	32,136	4.97	6,666	8.73	3,198	10.31	28,939	4.18	3,469	7.33
ind	17,341	41.1	3,846	55.2	1,799	54.55	15,543	39.56	2,048	55.74
ita	95,540	39.5	22,883	57.63	12,093	64.59	83,448	35.39	10,791	50.04
jpn	89,706	31.92	21,954	53.89	11,423	63.19	78,284	27.08	10,532	43.7
nno	11,670	47.37	3,217	59.49	1,751	64.94	9,920	45.66	1,467	56.95
nob	13,012	47.01	3,516	58.72	1,855	63.13	11,158	45.42	1,662	56.61
pol	69,365	36.29	16,353	58.22	8,398	65.55	60,968	30.91	7,956	49.82
por	120,069	46.11	26,044	62.48	13,486	65.64	106,584	42.82	12,559	58.84
slv	25,391	47.17	4,995	58.91	2,248	59.59	23,144	45.91	2,748	58.33
spa	114,452	38.68	28,609	46.34	15,517	46.46	98,936	37.78	13,093	46.22
swe	93,448	32.08	20,683	47.87	10,637	57.12	82,812	29.12	10,047	40.69
tha	15,660	27.77	3,602	50.11	1,784	53.62	13,877	23.85	1,819	46.56

Comments on the results

- Very different precision values for different languages:
  - The quality of the dictionary (not only the size)
  - The quality and completeness of the reference wordnet in OMW (not only the size (number of synset-variant pairs; the number of variants for each synset)

Language specific issues

- Vowel signs in Arabic and Hebrew
- Bulgarian: accents in Wiktionary (*алкохо́л*)
- Croatian: accents (bȍlēst, Európa, bȓdo). Wiktionary uses hbs (Croatian and Serbian words are mixed and some Serbian words are written in Cyrillic)
- Spanish: in OmegaWiki forms other than the lemma are included:
  - 14584110-n actínidos / 14584110-n actínido
  - 09605289-n adulto / 09605289-nadulta

Optimization of the weights

- 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/2:1:0:0:0:0:0;

  02675657-n/0:0:0:0:0:0:0

- Information: common lemmata: hyponym: hypernym: holonym: meronym: antonym: synonym

Experiments for Catalan

Def.	Rel.	Omegawiki	Wiktionary
0	1	70.01	**68.90**
1	0	70.95	66.15
1	1	71.03	66.14
1	5	**71.08**	66.15
5	1	70.98	66.14
1	10	71.06	65.90
10	1	70.98	66.14

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

….

Conclusions

- The disambiguation algorithm performs well
- There is room for improvement

Future work

- Experiment more complex disambiguation strategies
- Use other dictionaries and encyclopedias
- Run the algorithm for all languages in the resources
- Make contacts and agreements for the revision of the results
- Compare the results with Extended Open Multilingual Wordnet
- Pack the new algorithm and resources into WN-Toolkit

Thank you for your attention!

Antoni Oliver
aoliverg@uoc.edu