UNIGE Experiments on Robust Word Sense Disambiguation

Jacques Guyot, Gilles Falquet, Saïd Radhouani, Karim Benzineb

Centre universitaire d'informatique, University of Geneva

Route de Drize 7, 1227 Carouge

jacques.guyot, gilles.falquet, said.radhouani@unige.ch; karim@alpineblue.eu

Abstract

This task was meant to compare the results of two different retrieval techniques: the first one was based on the words found in documents and query texts; the second one was based on the senses (concepts) obtained by disambiguating the words in documents and queries. The underlying goal was to come up with a more precise knowledge about the possible improvements brought by word sense disambiguation (WSD) in the information retrieval process. The proposed task structure was interesting in that it drew up a clear separation between the actors (humans or computers): those who provide the corpus, those who disambiguate it, and those who query it. Thus it was possible to test the universality and the interoperability of the methods and algorithms involved.

Training and Testing Data

The document corpus was created by merging two collections of English documents : *LA Times 94* and *Glasgow Herald 9*5 (166'000 documents with 470'000 unique words and 55 million word occurrences). This corpus was processed with two different word sense disambiguation algorithms: UBC [ubc07] and NUS [nus07], resulting in two different sets. The disambiguation process replaced each occurrence of a term (composed by one or more words) by an XML element containing the term identifier, an extracted lemma, a part-of-speech (POS) tag (noun, verb, adjective…), the original word form (WF) and a list of senses together with their respective scores. The senses were represented by WordNet 1.6 synset identifiers. For instance, the word "discovery" could be replaced by:

```
<TERM ID="C041-19" LEMA="discovery" POS="NN">
   <WF>discovery</WF>
   <SYNSET SCORE="0.33057320536529267" CODE="04475449-N"/>
   ...
</TERM>
```

A training set of 150 queries (topics) was provided together with the expected results, as well as a testing set containing 160 queries. As usual, the queries included three parts: a title (T), a description (D) and a narrative (N). The English queries were processed with the UBC and NUS disambiguation algorithms, while the Spanish queries were disambiguated with the first sense heuristics (FSH), *i.e.* always choosing the first sense available.

Experiments

Indexer

To index the corpus, we chose the IDX-VLI indexer described in [gfb06] because it can gather a wealth of information (positions, etc.), it has built-in operators and it is remarkably fast. Still, we only used the basic version of that indexer, *i.e.* we did not use any relevance feedback mechanism, context description or any other sophisticated tool of that sort. We thus avoided interfering with the direct results of the experiment and we facilitated the result analysis.

Collection processing
We developed and tested several document processing strategies on the provided collections. Those strategies were applied to each <TERM> element within each document:

- NAT : Keep only the word form of each element (*i.e.* rebuild the original text)
- LEM : Keep only the lemma
- POS : Keep the lemma and the part-of-speech tag
- WSD : Keep only the synset with the best score \footnotemark[1].
- WSDL : Keep the best synset and the lemma.

During the indexing process the strategies were applied to all the terms, including numbers, except for the stop-words. Given the poor performance of the POS approach, we quickly gave up this option.

Topic processing
The same translators were applied to the queries, with an extended stop-word list including words such as *report*, *find*, etc. For each topic we derived three queries:

- T : Include only the title part
- TD : Include the title and description translated terms
- TDN : Include the title, description and narrative translated terms.

In order to come up with a reasonably good base line we tested several approaches to build a Boolean pre-filter from a given topic (results are the mean average precision (MAP) on T):

- OR (25.5%) : The logical OR of the terms (or lemmas)
- AND (15.8%) : The logical AND of the terms
- NEAR (15.2%) : The logical OR of all the pairs (t_i NEAR t_j) where t_i and t_j are the query terms
- AND-1 (23.6%): The logical OR of all the possible conjunctions of terms, except for the conjunction of all the terms.

The best results in terms of MAP were produced by the OR filtering, followed by the computation of a relevance score based on the Okapi BM25 weighting model (with default parameters). The test was carried out on the titles (T) of 150 training topics. More restrictive filtering schemes were tried out but did not perform any better, probably because of the relatively small size of the corpus.

Runs with word senses: For the disambiguation-based runs we tried out several other filtering schemes, including:

- OR (22.4%) : The logical OR of the best synset corresponding to a topic term
- AND (15.1%) : The logical AND of the best synset corresponding to a topic term
- NEAR (12.5%) : The logical OR of all the pairs (s_i NEAR s_j) where s_i and s_j are the best synsets corresponding to a topic term t_i and t_j
- AND-1 (18.8%): The logical OR of all the possible conjunctions of synsets, except for the conjunction of all the synsets
- HYPER (14.3%): The logical AND of each (s_i OR h_i) where s_i is the best synset corresponding to a topic term t_i and h_i is the direct hypernym of s_i in WordNet
- ORHYPER(18.43%): The logical OR of each (s_i OR h_i) where s_i is the best synset corresponding to a topic term t_i and h_i is the direct hypernym of s_i in WordNet.

However, none of these strategies performed any better than the basic OR filter on terms.

[1] This amounts to considering that the disambiguation algorithm is "perfect". Alternatively we could have added all the synsets with a score greater than a given threshold.

Result summary

The first table below shows the mean average precision (in percent) calculated on the training queries with different query processing options (disambiguation algorithm and part-of-topic selection) and different document processing options (disambiguation and translation). Of course, the <TERM> processing (NAT, LEM, WSD or WSDL) was always the same throughout the queries and the corpus for a given run.

The base line was the run with topic selection TDN and term selection LEM (*i.e.* the whole topic with stemming).

The second table shows the results of the testing queries, which are slightly better than those of the training queries (maybe the testing queries were somewhat easier).

The base line (LEM) for the Spanish queries was created by automatically translating the queries from Spanish into English.

The tests on the NUS corpus produced better results than those on the UBC one. Therefore most of the runs were performed on the NUS corpus, while the UBC corpus would be used to test the interoperability of the disambiguation processes.

avg precision on **TRAINING** requests (OR strategy)			Document processing				
			Base Line		NUS		UBC
			NAT	LEM	WSD	WSD+LEM	WSD
Request	NONE	T	25.2%	27.0%			
		TDN		31.9%			
	NUS	T			22.4%	26.0%	
		TDN			28.8%	32.5%	24.9%
	UBC	T			22.6%		
		TDN			22.4%		25.4%
	ESP	T			4.0%		
		TDN			6.6%		6.2%

avg precision on **TESTING** requests (OR strategy)			Document processing			
			Base Line	NUS		UBC
			LEM	WSD	WSD+LEM	WSD
Request	NONE	T	30.64%			
		TD	36.64%			
		TDN	39.17%			
	NUS	T		21.20%		
		TD		29.34%		
		TDN		32.69%	38.14%	
	UBC	TDN		29.62%		
	Trans. ESP	T	30.36%			
	FSH-ESP	T		8.46%		
		TDN		9.70%		

Findings and Discussion

In the tables above we note the following facts:

- Using the D and N parts-of-topics increases the precision in all cases (with and without WSD). This is probably due to the ranking method which benefits from the additional terms provided by D and N.
- On the test run with UBC disambiguation, the senses alone (WSD) decrease the MAP: -4.6% on T queries and -3.1% on TDN. On training requests, adding the lemmas to the senses (WSDL) slightly improves the MAP (+0.6%). This is the only case where disambiguation brings any improvement.
- Using different disambiguation algorithms for the queries and the documents noticeably decreases the results. This should not happen if the algorithms were perfect. It shows that disambiguation acts as a kind of encoding process on the words, and obviously the best results are obtained when the same encoding, producing the same mistakes, is applied to both queries and documents. Thus, at this stage, the disambiguation algorithms are not interoperable.

We carefully analyzed about 50 queries to better understand what happened with the disambiguation process. For instance, the query with the title "*El Niño and the weather*" was disambiguated as follows (NUS):

- **El** was understood as the abbreviation *el.* of *elevation*
- **Niño** was understood as the abbreviation *Ni* of *nicke*l, probably because the parser failed on the non-ASCII character *ñ*
- **weather** was correctly understood as the *weather* concept.

Although the disambiguation was incorrect, WSD was as good as LEM because the "encoding" was the same in the collection and in the query and there were few or no documents about nickel that could have brought up noise.

More generally, when the WSD results were better than the LEM ones, it was not due to semantic processing but to contingencies. For instance, the query title "*Teenage Suicides*" had a better score with WSD because *teenage* was not recognized! Thus the query became *suicides*, which is narrower than *teenage OR suicide* and, on this corpus, avoids retrieving a large amount of irrelevant documents about teenagers.

A few items of the test run are commented in Appendix A.

The poor performance on Spanish queries is due to 1) the above-mentioned lack of interoperability between the different WSD algorithms, and 2) the low quality of the Spanish WSD itself.

This can be illustrated with some examples:

On Question 41: "Pesticide in baby food" is translated by "Pesticidas en alimentos para bebes" and is then converted into the FOOD and DRINK (verb) concepts because *bebes* is a conjugated form of *beber*, which is the Spanish verb for drink.

On Question 43: "El Niño and the weather" is translated by "El Niño y el tiempo" and is then converted into the CHILD and TIME concepts because *Niño* is the Spanish noun for child and *tiempo* is an ambiguous word meaning both time and weather.

Given those difficulties, outstanding results could not be expected.

Looking back on the questions and results, it can be noted that 1'793 documents were retrieved out of the 2'052 relevant ones, *i.e.* almost 90% of them. The core issue is to sort out the documents so as to reject those whose content does not match users' expectations.

A closer look at our results on the Training corpus showed that we got a pretty good performance on some of the requests. This does not mean that our search engine understood the said requests correctly; it is simply due to the fact that the corpus included only good matches for those requests, so it was almost impossible to find wrong answers.

For instance, on Question 50 about "the Revolt in Chiapas", we retrieved 106 documents out of the 107 relevant ones with an average precision of 87%. This is due to the fact that in the corpus, the Chiapas are only known for their revolt (in fact if we google the word "Chiapas" a good proportion of the results are currently about the Chiapas rebellion).

On the other hand, on Question 59: "Computer Viruses", our search engine retrieved 1 document on 1 with an average precision of 0.3%. This is because the 300 documents retrieved before the one we were looking for were indeed about viruses and computers, but did not mention any virus name or damage as was requested.

Therefore term disambiguation does not help the search engine to understand what kind of documents are expected. A question such as the one above requires the text to be read and understood in order to decide whether it is actually a correct match.

Conclusion

Intuitively, Word Sense Disambiguation should improve the quality of information retrieval systems. However, as already observed in previous experiments, this is only true in some specific situations, for instance when the disambiguation process is almost perfect, or in limited domains. The observations presented here seem to support this statement. We propose two types of explanations:

1. When a query is large enough (more than one or two words), the probability that a document containing these words uses them with a meaning different from the intended one is very low. For instance, it is unlikely that a document containing *mouse*, *cheese* and *cat* is in fact about a computer mouse. This probably makes WSD useless in many situations. Such a request is similar in nature to the narrative-based tests. On the other hand, the WSD approach could make more sense when requests include only one or two words (which is the most frequent case in standard searches).

2. WSD is a very partial semantic analysis which is insufficient to really understand the queries. For instance, consider the query "*Computer Viruses*" whose narrative is "Relevant documents should mention the name of the computer virus, and possibly the damage it does". To find relevant documents, a system must recognize phrases which contain virus names ("the XX virus", "the virus named XX", "the virus known as XX", etc.). It should also recognize phrases describing damages ("XX erases the hard disk", "XX causes system crashes" but not "XX propagates through mail messages"). These tasks are very difficult to perform and they are far beyond the scope of WSD. Moreover, they require specific domain knowledge, as shown in [rf06] .

The modifications brought to our stop-word lists showed that our search engine is more sensitive to various adjustments of its internal parameters than to the use of a WSD system. Indeed, when we ran a new series of tests with English-only stop words (which eliminated some terms in the requests, such as "eu" and "un"), our new score for the LEM-TDN (which was our best result in this task) increased from 39.17% to 39.63%.

Finally, as we argued in [grf05], conceptual indexing is a promising approach for language-independent indexing and retrieval systems. Although an efficient WSD is essential to create good conceptual indexes, we showed in [grf05] that ambiguous indexes (with several concepts for some terms) are often sufficient to reach a good multilingual retrieval performance, for the reasons mentioned above.

Bibliography

[gfb06] Guyot, J., Falquet, G., Benzineb, K. (2006) Construire un moteur d'indexation. Technique et science informatique (TSI), Hermes, Paris.

[grf05] Guyot, J., Radhouani, S., Falquet, G. (2005) Conceptual Indexing for Multilingual Information Retrieval. In Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers. C. Peters, et al. (Eds.). Lecture Notes in Computer Science, Vol. 4022, Springer.

[nus07] Chan, Y. S., Hwee T., Zhong, Z. (2007) NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks. Proc. of the 4th International Workshop on Semantic Evaluations (SemEval 2007). Prague, Czech Republic. pp 253--256.

[rf06] Radhouani, S., Falquet, G. (2006) Using External Knowledge to Solve Multi-Dimensional Queries, in Proc. 13th Intl Conf. on Concurrent Engineering Research and Applications (CE 2006), Antibes, Sept. 2006. IOS Press.

[ubc07] Agirre, E., Lopez de Lacalle, O. (2007) UBC-ALM: Decombining k-NN with SVD for WSD. Proc. of the 4th International Workshop on Semantic Evaluations (SemEval 2007). Prague, Czech Republic. pp 341--345.

Appendix A: Comparing the LEM and the WSD approaches

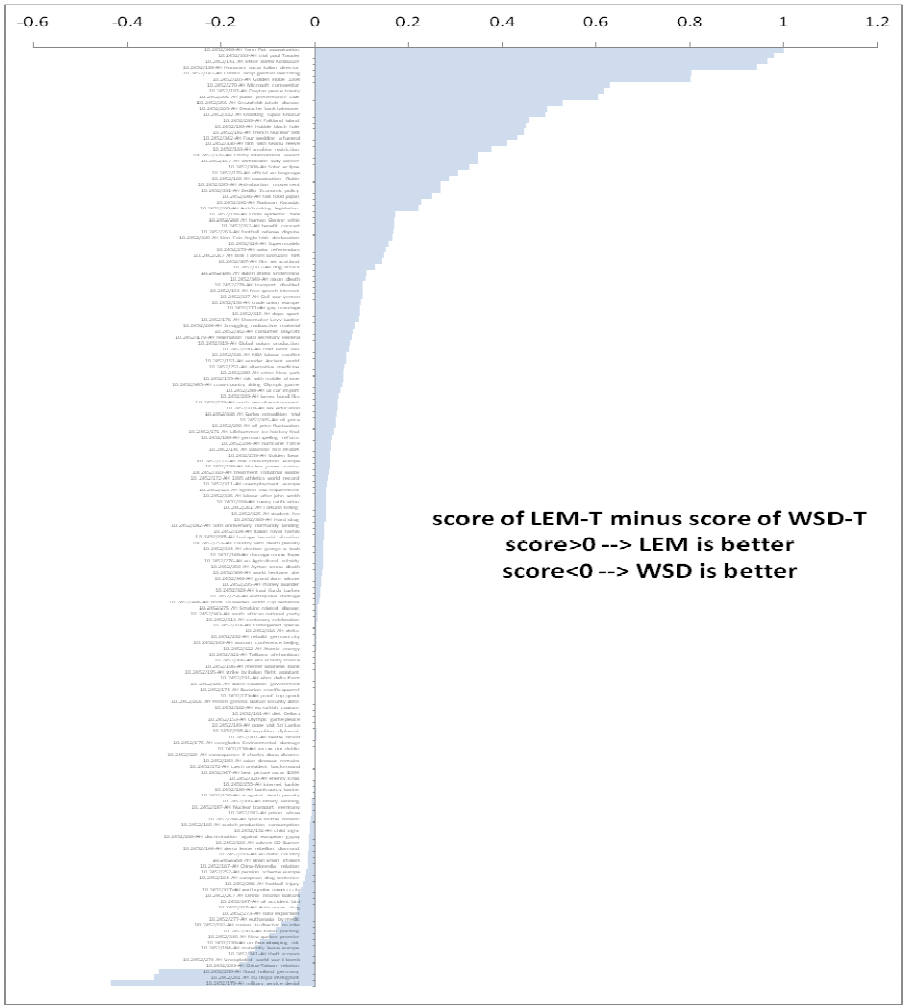
			LEM-T		WSD-T		
Query	Title	Relevant Doc.	Retrieved Doc.	Avg. Prec.	Retrieved Doc.	Avg. Prec.	LEM-WSD
178	military service denial	4	3	0.1271	3	0.5625	-0.4354
291	eu Illegal immigrant	30	23	0.0206	25	0.3642	-0.3436
200	flood holland germany	9	9	0.3712	9	0.7034	-0.3322
293	China-Taiwan relation	34	8	0.0049	26	0.1658	-0.1609
274	Unexploded world war ii bomb	16	12	0.1767	13	0.3215	-0.1448
341	theft scream	6	5	0.2594	6	0.3997	-0.1403
184	maternity leave europe	9	8	0.332	9	0.4597	-0.1277
299	un Peacekeeping risk	76	39	0.0532	73	0.1713	-0.1181
340	New quebec premier	5	5	0.1975	5	0.2942	-0.0967
303	italian painting	14	12	0.421	12	0.5029	-0.0819
192	russian tv director murder	6	6	0.2156	6	0.2874	-0.0718
277	euthanasia by medic	26	12	0.0798	15	0.1272	-0.0474
273	nato expansion	83	73	0.5305	77	0.5691	-0.0386
317	Anti-cancer drug	30	9	0.1072	22	0.1458	-0.0386
147	oil accident bird	51	49	0.5885	47	0.626	-0.0375
257	Ethnic cleanse balkans	63	41	0.0918	43	0.1279	-0.0361
327	earthquake mexico city	4	4	0.4512	4	0.4821	-0.0309
286	football injury	12	7	0.0426	7	0.0692	-0.0266
164	european drug sentence	27	25	0.1569	24	0.18	-0.0231
252	pension scheme europe	29	29	0.2982	29	0.3183	-0.0201
167	China-Mongolia relation	5	1	0.001	2	0.0179	-0.0169
258	Brain-Drain impact	4	0	0	4	0.0162	-0.0162
193	eu baltic country	7	5	0.2411	4	0.2571	-0.016
144	sierra leone rebellion	3	3	0.2667	3	0.281	-0.0143
	diamond						
169	advent CD-Burner	6	1	0.0005	4	0.0148	-0.0143
266	discrimination against european gypsy	4	4	0.0676	4	0.0813	-0.0137
152	child right	11	10	0.1494	11	0.1619	-0.0125
160	scotch production consumption	2	1	0.0152	2	0.0262	-0.011
284	space shuttle mission	38	38	0.3865	38	0.3962	-0.0097
282	prison abuse	31	8	0.0082	9	0.0158	-0.0076
187	Nuclear transport germany	1	1	0.0333	1	0.04	-0.0067
300	lottery winning	42	42	0.2231	38	0.2297	-0.0066
150	ai against death penalty	10	5	0.0029	7	0.0065	-0.0036
180	bankruptcy baring	40	40	0.4997	40	0.5029	-0.0032
255	internet junkie	4	4	0.1388	4	0.1416	-0.0028
320	energy crisis	16	6	0.0086	6	0.011	-0.0024
347	best picture oscar 1994	9	8	0.0481	8	0.0502	-0.0021
272	czech president background	14	6	0.02	6	0.0221	-0.0021
183	asian dinosaur remains	4	4	0.1754	4	0.1774	-0.002
329	consequence if charles diana divorce	17	17	0.4143	17	0.4161	-0.0018
158	soccer riot dublin	34	20	0.2092	20	0.2102	-0.001
175	everglades Environmental damage	7	7	0.6606	7	0.6611	-0.0005
301	Nestlé brand	16	5	0.0096	5	0.0099	-0.0003
297	expulsion diplomat	12	12	0.3603	12	0.3604	-1E-04
149	pope visit Sri Lanka	0	0	0	0	0	0
153	Olympic game peace	1	0	0	0	0	0
161	diet Celiacs	0	0	0	0	0	0
162	eu turkish custom	1	1	0.2	1	0.2	0
166	french general Balkan security zone	0	0	0	0	0	0

173	proof top quark	2	2	1	2	1	0
174	Bavarian crucifix quarrel	2	2	1	2	1	0
186	dutch coalition government	0	0	0	0	0	0
191	ebro delta farm	0	0	0	0	0	0
195	strike by italian flight assistant	0	0	0	0	0	0
196	merger japanese bank	1	1	1	1	1	0
306	eta activity france	1	1	1	1	1	0
321	Talibans afghanistan	0	0	0	0	0	0
322	Atomic energy	14	11	0.0295	8	0.0287	0.0008
143	woman conference beijing	39	39	0.845	39	0.844	0.001
292	rebuild german city	4	3	0.1314	2	0.1293	0.0021
316	strike	118	103	0.2022	105	0.1992	0.003
314	Endangered specie	18	13	0.0345	15	0.0313	0.0032
313	centenary celebration	20	8	0.0115	8	0.007	0.0045
343	south african national party	1	1	0.0175	1	0.013	0.0045
275	Smoking-related disease	66	23	0.0266	11	0.0211	0.0055
344	brazil vs sweden world cup semifinal	33	33	0.546	33	0.5379	0.0081
254	earthquake damage	229	174	0.2636	169	0.2533	0.0103
328	iraqi Kurds turkey	23	23	0.439	22	0.4283	0.0107
295	money launder	51	27	0.2699	24	0.2576	0.0123
346	grand slam winner	11	10	0.1011	9	0.0872	0.0139
304	world heritage site	15	9	0.2383	8	0.2238	0.0145
350	Ayrton senna death	10	10	0.6583	10	0.6417	0.0166
276	eu Agricultural subsidy	59	57	0.3416	50	0.3249	0.0167
148	damage ozone layer	6	5	0.1836	1	0.1667	0.0169
334	election george w bush	5	4	0.0675	4	0.0499	0.0176
253	country with death penalty	190	133	0.3116	140	0.2934	0.0182
287	hostage terrorist situation	49	47	0.2731	46	0.254	0.0191

194	italian royal family	1	1	0.0213	1	0.0012	0.0201
182	50th anniversary normandy landing	4	4	0.0399	4	0.0196	0.0203
309	Hard drug	10	9	0.0225	0	0	0.0225
325	student fee	61	58	0.454	58	0.4313	0.0227
261	Fortune-telling	34	2	0.0231	1	0.0003	0.0228
269	treaty ratification	9	8	0.128	8	0.1051	0.0229
335	labour after john smith	15	12	0.0827	13	0.0593	0.0234
323	tighten visa requirement	26	14	0.0559	12	0.0322	0.0237
311	unemployment europe	54	49	0.1991	47	0.1736	0.0255
172	1995 athletics world record	7	6	0.0358	4	0.0092	0.0266
310	treatment Industrial waste	63	39	0.0679	31	0.0382	0.0297
298	Nuclear power station	30	25	0.1033	21	0.0721	0.0312
177	milk consumption europe	13	13	0.3448	13	0.3135	0.0313
259	Golden bear	3	2	0.0341	1	0.0015	0.0326
145	japanese rice import	38	34	0.4846	33	0.4516	0.033
294	hurricane force	27	27	0.4374	26	0.4027	0.0347
188	german spelling reform	1	1	0.0385	1	0.0031	0.0354
171	Lillehammer ice hockey final	19	14	0.0407	1	0.0001	0.0406
290	oil price fluctuation	49	49	0.2299	42	0.1864	0.0435
305	oil price	61	60	0.276	46	0.2323	0.0437
338	Carlos extradition trial	10	8	0.0459	0	0	0.0459
318	sex education	9	9	0.1756	9	0.1283	0.0473
159	north sea oil environment	54	44	0.3161	49	0.268	0.0481
283	james bond film	38	33	0.4357	34	0.3861	0.0496
288	us car import	76	48	0.1248	41	0.0697	0.0551
345	cross-country skiing Olympic game	9	8	0.0644	3	0.0062	0.0582
155	risk with mobile phone	3	3	0.1104	3	0.0505	0.0599

280	crime New york	25	21	0.1445	19	0.083	0.0615
251	alternative medicine	65	33	0.2392	15	0.1762	0.063
151	wonder Ancient world	12	9	0.444	9	0.3772	0.0668
336	NBA labour conflict	17	10	0.0679	0	0	0.0679
190	child labor asia	9	6	0.4935	7	0.4213	0.0722
319	Global opium production	19	14	0.2615	10	0.1865	0.075
179	resignation nato secretary general	22	21	0.3122	20	0.2337	0.0785
302	consumer boycott	17	17	0.3806	15	0.2981	0.0825
264	Smuggling radioactive material	22	21	0.2916	18	0.2053	0.0863
176	Shoemaker-Levy jupiter	39	39	0.8747	39	0.7809	0.0938
315	dope sport	59	36	0.2969	30	0.2016	0.0953
271	gay marriage	24	23	0.4388	23	0.3425	0.0963
156	trade union europe	22	19	0.2132	19	0.115	0.0982
337	Civil war yemen	20	20	0.8316	19	0.7301	0.1015
154	free speech internet	21	19	0.1849	17	0.0829	0.102
278	transport disabled	21	15	0.1477	16	0.0453	0.1024
349	nixon death	27	27	0.2799	27	0.1724	0.1075
185	dutch photo Srebrenica	1	1	0.1111	0	0	0.1111
312	dog attack	31	31	0.6389	31	0.5105	0.1284
307	film set scotland	77	70	0.3375	69	0.1967	0.1408
267	best Foreign language film	21	17	0.1627	4	0.017	0.1457
279	swiss referendum	4	4	0.4054	4	0.2564	0.149
324	Supermodels	74	18	0.156	0	0	0.156
339	Sinn Fein Anglo-Irish declaration	19	19	0.6059	19	0.4419	0.164
263	football referee dispute	48	46	0.1784	7	0.0119	0.1665
262	benefit concert	85	67	0.3426	31	0.1738	0.1688
268	human Cloning ethic	3	2	0.1778	2	0.009	0.1688
199	Ebola epidemic zaire	10	10	0.4445	10	0.2741	0.1704
260	Anti-Smoking legislation	63	48	0.2339	14	0.0132	0.2207
281	Radovan Karadzic	42	42	0.2256	0	0	0.2256
146	fast food japan	2	1	0.5	1	0.25	0.25

331	Zedillo Economic policy	17	16	0.2704	7	0.003	0.2674
285	Anti-abortion movement	69	62	0.6232	67	0.3547	0.2685
168	assassination Rabin	18	18	0.3055	12	0.019	0.2865
170	official eu language	1	1	0.3333	1	0.0278	0.3055
308	Solar eclipse	11	11	0.7414	10	0.4139	0.3275
157	Wimbledon lady winner	139	109	0.3473	4	0.0001	0.3472
326	Emmy international award	4	4	0.433	2	0.0857	0.3473
163	smoking restriction	122	120	0.6032	89	0.2279	0.3753
330	film with Keanu reeve	51	50	0.6538	48	0.2454	0.4084
342	Four wedding a funeral	96	90	0.6021	77	0.171	0.4311
181	french Nuclear test	92	91	0.8384	84	0.3933	0.4451
189	Hubble black hole	7	7	0.8828	7	0.4316	0.4512
289	Falkland island	18	17	0.4692	13	0.0137	0.4555
332	Shooting Tupac Shakur	8	8	0.4924	0	0	0.4924
265	Deutsche bank takeover	6	6	0.5337	4	0.0373	0.4964
256	Creutzfeldt-Jakob disease	34	30	0.6114	28	0.0838	0.5276
296	public performance Liszt	24	24	0.6043	1	0	0.6043
197	Dayton peace treaty	50	50	0.6219	7	0.0051	0.6168
270	Microsoft competitor	57	55	0.6301	11	0.0028	0.6273
165	Golden globe 1994	1	1	1	1	0.2	0.8
142	Christo wrap german Reichstag	8	8	0.9472	6	0.1443	0.8029
198	Honorary oscar italian director	1	1	1	1	0.0588	0.9412
141	letter bomb Kiesbauer	1	1	1	1	0.037	0.963
333	trial paul Touvier	5	5	1	5	0.0222	0.9778
348	Yann Piat assassination	2	2	1	0	0	1


-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
score of LEM-T minus score of WSD-T
score>0 --> LEM is better
score<0 --> WSD is better

Appendix B: Analysis of some requests which performed better with WSD

Case 1)

Query 178	military OR service OR denial	Av. Prec.: 12.71%
Query 178	(WSD06091176-n OR WSD05382699-n)	Av. Prec.: 56.25%

Associated Concepts

WSD06091176-n military_service WSD06092672-n

WSD05382699-n denial WSD05045355-n

Interpretation

Clearly here the WSD process improved the results because the search engine looked for the *military_service* concept instead of looking for *military* OR *service*. This is the ideal situation and we could have expected it to be the most frequent case, but it was not, as illustrated below.

Case 2)

Query 291	eu OR Illegal OR immigrant	Av. Prec.: 2.06%
Query 291	(WSD10485926-n OR WSD01346039-a OR WSD07334599-n)	Av.Prec.: 36.42%

Associated Concepts

WSD10485926-n europium WSD10476248-n

WSD01346039-a illegal WSD01346510-a

WSD07334599-n immigrant WSD07408670-n

Interpretation

What happened here is that *eu* appears in the French list of stop-words, so the request becomes less discriminatory since it is only based on the two other words *illegal* and *immigrant*. When the stop-word list is restricted to English words only, the new average precision of the lemma-based search is 35.94%, *i.e.* almost as good as the WSD request.

Case 3)

Query 200	flood OR holland OR germany	Av. Prec.: 37.12%
Query 200	(WSD00450672-n OR WSD06536741-n OR WSD06442182-n)	Av. Prec.: 70.34%

Associated Concepts

WSD00450672-n implosion_therapy

WSD00449552-n S: (n) implosion therapy, flooding (a technique used in behavior therapy; client is flooded with experiences of a particular kind until becoming either averse to them or numbed to them)

WSD06536741-n Netherlands WSD06401678-n

WSD06442182-n Germany WSD06401678-n

Interpretation

Two explanations may be proposed in this case: *flooding* was wrongly encoded but it still specifies the concept better than the *flood* lemma; and *Holland* was encoded as *Netherlands*, so the documents were badly classified because the search engine did not know that the two words are synonyms.

Case 4)

Query 293	China-Taiwan OR relation	Av. Prec.: 0.0049
Query 293	(WSD06417803-n OR WSD00018916-n)	Av. Prec.: 0.1658

Associated Concepts

06417803-n China WSD06404073-n

WSD09924967-n causality WSD00018916-n

Interpretation

When searching for the *China-Taiwan* concept using the lemmas, the engine found only 9 documents out of the expected 34, while the WSD process retrieved 26 documents. It is noted here that the encoding process made two mistakes: *China-Taiwan* was encoded as *China* only, and *relation* was encoded as *causality*. Yet the WSD average precision is still better because the same encoding mistakes were made systematically across the corpus and queries.