Approximating hierarchy-based similarity for WordNet nominal synsets using Topic Signatures

Eneko Agirre
Enrique Alfonseca
Oier López de Lacalle

2^{nd} Global WordNet Conference — Brno, January 20–23, 2004

Index

Introduction (I)

- A **topic signature** is a topical vector relevant to a word sense
  $\Rightarrow$ it contains terms which tend to appear in its context (but not with other senses).
- It is possible to extend WordNet synsets with topic signatures automatically.

Introduction (I)

- A **topic signature** is a topical vector relevant to a word sense
  $\Rightarrow$ it contains terms which tend to appear in its context (but not with other senses).
- It is possible to extend WordNet synsets with topic signatures automatically.

Applications:

- Word-sense disambiguation (weak).
- Clustering of word senses.
- Populating automatically WordNet with new concepts.
- In general, as a substitute for other similarity metrics.

Introduction (II)

church, Christianity:
church(1177.83) catholic(700.28) orthodox(462.17) roman(353.04)
religion(252.61) byzantine(229.15) protestant(214.35) rome(212.15)
western(169.71) established(161.26) coptic(148.83) jewish(146.82)
...

Introduction (II)

church, Christianity:
church(1177.83) catholic(700.28) orthodox(462.17) roman(353.04) religion(252.61) byzantine(229.15) protestant(214.35) rome(212.15) western(169.71) established(161.26) coptic(148.83) jewish(146.82) ...

church, church_building:
house(1733.29) worship(1079.19) building(620.77) mosque(529.07) place(507.32) synagogue(428.20) god(408.52) kirk(368.82) build(93.17) construction(47.62) street(47.18) nation(41.16) road(40.12) congregation(39.74) muslim(37.17) list(34.19) construct(31.74) ...

Introduction (II)

church, Christianity:
church(1177.83) catholic(700.28) orthodox(462.17) roman(353.04) religion(252.61) byzantine(229.15) protestant(214.35) rome(212.15) western(169.71) established(161.26) coptic(148.83) jewish(146.82) ...

church, church_building:
house(1733.29) worship(1079.19) building(620.77) mosque(529.07) place(507.32) synagogue(428.20) god(408.52) kirk(368.82) build(93.17) construction(47.62) street(47.18) nation(41.16) road(40.12) congregation(39.74) muslim(37.17) list(34.19) construct(31.74) ...

church, church_service:
service(5225.65) chapel(1058.77) divine(718.75) prayer(543.96) hold(288.08 cemetery(284.48) meeting(271.04) funeral(266.05) sunday(256.46) morning(169.38) attend(143.64) pm(133.56) meet(115.86) conduct(98.96) wednesday(90.13) religious(89.19) evening(75.01) day(74.45) friday(73.17) eve(70.01) monday(67.96)...

Introduction (III)

The purpose of this work is

- To compare similarity measures for WordNet concepts based on topic signatures against other metrics based on WordNet.

  $\Rightarrow$ It will be possible to apply these kinds of measures to unknown concepts.
- To study different ways of acquiring and modelling the signatures.

Constructing the topic signatures (I)

1. Search the Internet to collect texts related to that sense
   ⇒ Use WordNet 1.7 in building the queries
2. Store a collection of documents for each sense.
3. Extract the words and frequencies from each collection.
4. Apply a formula to find the words with a distinctive frequency for a collection.
5. Store them in the **topic signature**.

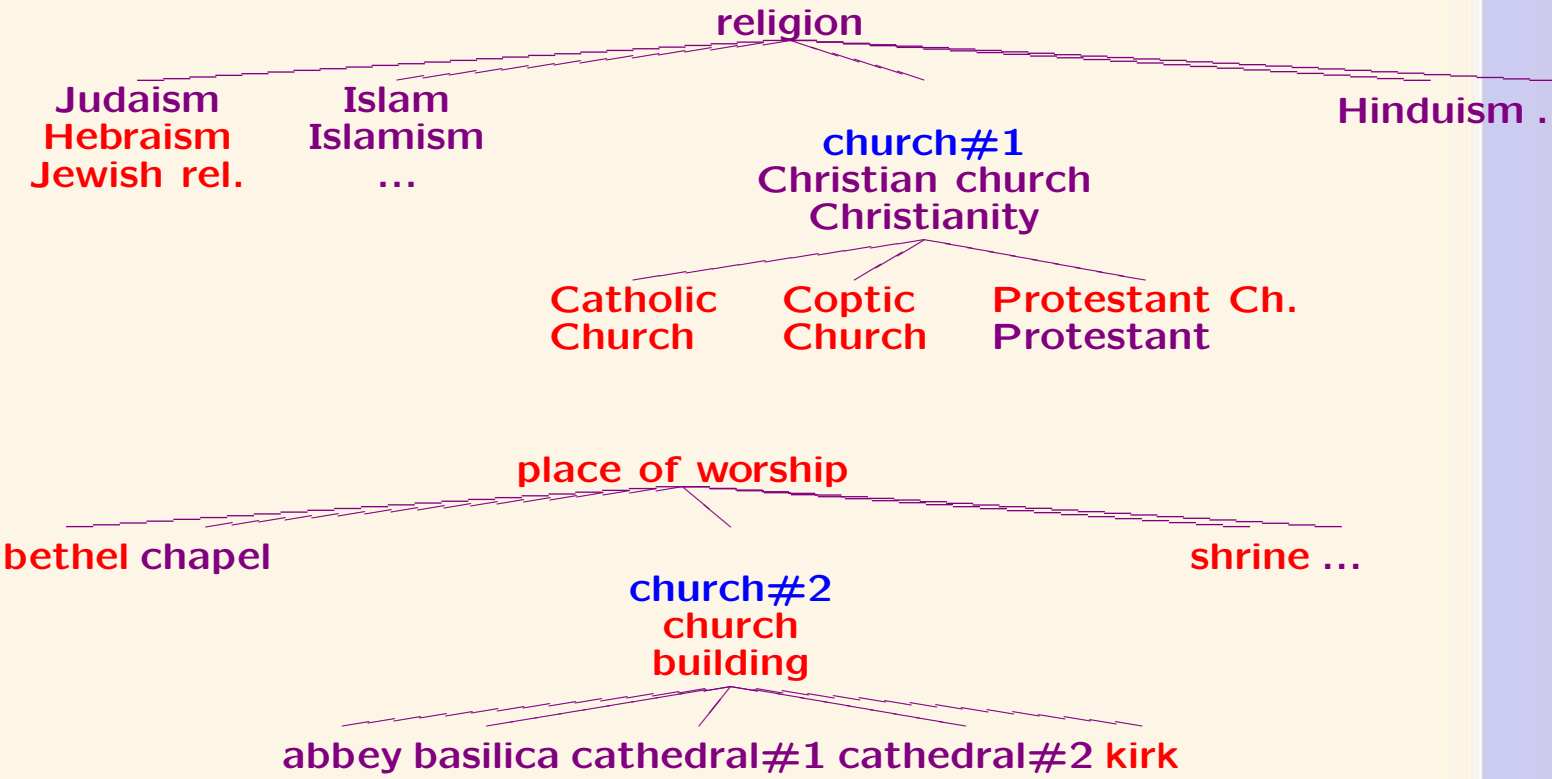
Acquiring examples (I)

We checked two possibilities for using WordNet to build the queries:

- Use all relatives of the word sense.
- Use only the monosemous relatives.

Acquiring examples (II)

Example:



Acquiring examples (III)

Implementation:

Queries	Search engine	Documents
Monosemous	Google	1000 snippets
All relatives	Altavista	100 documents

Representing context (I)

- Vectorial representation of context (*bag of words*)
- Each word sense is represented as a vector of $\mathcal{V}$ dimensions, where the i^{th} element contains the frequency of the i^{th} word in $\mathcal{V}$.
- All the words in the documents/snippets are stemmed.

Representing context (II)

(1) a. The oldest preserved church building of the Prague Castle.

b. I remember him in the mornings sweeping the street and church stairs nothing fatheaded about him.

c. There were several other church fires, during his days, but "never major."

d. He was appointed assistant director of our capital campaign to raise funds to renvate the church building.

Word	Freq	Word	Freq
building	2	old	1
preserve	1	Prague	1
remember	1	morning	1
sweep	1	street	1
stair	1	fatheaded	1
fire	1	day	1
appoint	1	assistant	1
director	1	capital	1
campaign	1	raise	1
fund	1	renvate	1
major	1		

Weighting

Once we have the vectors corresponding to each word sense, we use a function to calculate the relevance of each term in each vector:

- The χ^2 function.
- Mutual Information.
- The t-score.
- Two versions of tf·idf:
  1. $\frac{tf_t}{max_t tf_t} \times \log \frac{N}{df_t}$
  2. $(0.5 + \frac{0.5 \times tf_t}{max_t tf_t}) \log \frac{N}{df_t}$

Filtering (I)

- Rare words that happen to be in one context by chance usually receive a large weight:
  $\Rightarrow$ proper nouns, mispelled words...
- A large corpus of English is used to filter out these words (the BNC).

Filtering (II)

Procedure

Filter(w**,** $signatures$**,** $corpus$**)**:

1. Collect all the contexts of w in the $corpus$ in a vector $\mathcal{C}$.
2. For each signature s_i (corresponding to one of w's senses):
   - Remove all the words from s_i that do not appear in $\mathcal{C}$.

Filtering (III)

Example:

Word	Freq	Word	Freq
building	2	old	1
preserve	1	Prague	1
remember	1	morning	1
sweep	1	street	1
stair	1	fatheaded	1
fire	1	day	1
appoint	1	assistant	1
director	1	capital	1
campaign	1	raise	1
fund	1	renvate	1
major	1		

Hierarchical-based similarity

The following similarity metrics between two word senses, based on the structure of WordNet, have been considered:

- Resnik's distance metric (based on the Information Content of the synset; probabilities obtained from Semcor).
- The inverse of the minimal number of hyperonymy links between the two synsets (*conceptual distance*).
- The coarse-grained distances used in the WSD exercise Senseval-2.

A similarity based on the topic signatures

Distance between topic signatures:

If we have two word senses, w_1 and w_2, with their respective topic signatures s_1 and s_2, two possible distance metrics between them are:

(a) Cosine:

$$d_1(s_1, s_2) = cosine(s_1, s_2)$$

(b) Euclidean:

$$d_2(s_1, s_2) = \sqrt{\sum_i (s_{1i} - s_{2i})^2}$$

Experiment (I)

Evaluation:

- Done with 16 nouns from the Senseval-2 evaluation.

art	authority	bar	bum
chair	channel	child	church
circuit	day	dike	facility
fatigue	feeling	grip	hearth

Experiment (II)

In building the topic signatures, the following **parameters** have been varied:

- Ways for constructing the queries (*monosemoous vs. all relatives*)
- Weight function (χ^2, *tf·idf*, *MI* or *t-score*).
- Filtering (*with* or *without*).
- Similarity metric (*cosine* or *euclidean*).
- Gold-standard metrics (*Resnik*, *link*, *coarse-grained senses*)

Results (I) – Monosemous queries

		Chi2		$Tf{\cdot}idf_1$		$Tf{\cdot}idf_2$		MI		t-score	
		No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
GoldStd.	Metric										
Sensev	Euc.	0.14	0.12	0.25	0.23	0.19	0.08	0.3	0.33	0.33	0.29
	cos	0.22	0.21	0.38	0.47	0.34	0.37	0.39	0.17	0.2	
Resnik	Euc.	0.31	0.28	0.35	0.44	0.28	0.39	0.56	0.56	0.55	0.51
	cos	0.39	0.38	0.35	0.26	0.35	0.37	0.52	0.49	0.31	0.35
dist	Euc.	0.63	0.61	0.63	0.7	0.48	0.51	0.81	0.87	0.87	0.8
	cos	0.47	0.45	0.65	0.6	0.69	0.72	0.88	0.87	0.61	0.62

- The conceptual distance metric, easier to approximate.

- Best results: MI, t-score.

Results (II) – All-relatives queries

		Chi2		Tf·idf$_1$		Tf·idf$_2$		MI		t-score	
		No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
GoldStd.	Metric										
Sensev	Euc.	0.17	0.16	0.18	0.18	0.18	0.18	0.33	0.33	0.35	0.32
	cos	0.33	0.3	0.33	0.39	0.34	0.39	0.32	0.34	0.03	0.04
Resnik	Euc.	0.38	0.37	0.3	0.3	0.3	0.3	0.48	0.49	0.49	0.47
	cos	0.44	0.28	0.47	0.46	0.51	0.48	0.65	0.61	0.42	0.3
dist	Euc.	0.65	0.65	0.57	0.57	0.57	0.57	0.82	0.84	0.85	0.82
	cos	0.49	0.43	0.62	0.84	0.44	0.43	0.81	0.84	0.44	0.43

- Again, the conceptual distance metric was easier to approximate.

- Best results: MI, t-score, tf·idf.

Conclusions (I)

Accuracy in approximating distances:

- The conceptual distance could be accurately approximated with the topic signatures (0.88).
- Resnik's metric, on the other hand, has not been so easy to approximate. In particular, two words in separate taxonomies had a similarity 0 with Resnik's metric *(e.g. the 3 senses of church)*.
- Finally, the coarse-grained sense, being binary, has proved the hardest to approximate.

Conclusions (II)

Accuracy with different parameters:

- Monosemous relatives produces better results.
- Filtering does not improve the similarity!
- MI and t-score produced better results!

Future work

- Experiments on a larger set of concepts, or from the same sub-taxonomy in WordNet.
- Compare to yet more similarity measures using WordNet.
- Repeat the experiment with signatures that model syntactic dependencies between the concepts.
- Explore further parameters in topic signature construction.
- Evaluate on an application (e.g. Ontology population).