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

Eneko Agirre[1], Enrique Alfonseca[2], and Oier Lopez de Lacalle[1]

[1] University of the Basque Country, Donostia 20.080, Spain,
{eneko,jibloleo}@si.ehu.es,
WWW home page: http://ixa.si.ehu.es
[2] Universidad Autonoma de Madrid,
enrique.alfonseca@ii.uam.es,
WWW home page: http://www.ii.uam.es/~ealfon

Abstract. Topic signatures are context vectors built for concepts. They can be automatically acquired for any concept hierarchy using simple methods. This paper explores the correlation between a distributional-based semantic similarity based on topic signatures and several hierarchy-based similarities. We show that topic signatures can be used to approximate link distance in WordNet (0.88 correlation), which allows for various applications, e.g. classifying new concepts in existing hierarchies. We have evaluated two methods for building topic sigantures (monosemous relatives vs. all relatives) and explore a number of different parameters for both methods.

1 Introduction

Knowledge acquisition is a long-standing problem in both Artificial Intelligence and Natural Language Processing (NLP). Huge efforts and investments have been made to manually build repositories with semantic and pragmatic knowledge but with unclear results. Complementary to this, methods to induce and enrich existing repositories have been explored (see [9] for a recent review).

In previous work we have shown that it is possible to enrich WordNet synsets [7] with topic signatures. Topic signatures try to associate a topical vector to each word sense. The dimensions of this topical vector are the words in the vocabulary and the weights try to capture the relatedness of the words to the the target word sense. In other words, each word sense is associated with a set of related words with associated weights. Figure 1 shows sample topic signatures for the word senses of church. Several of the topic signatures used in this paper can be found in http://ixa3.si.ehu.es/cgi-bin/signatureak/signaturecgi.cgi in its full version.

Topic signatures for words have been successfully used in summarisation tasks [8]. Regarding topic signatures for word senses, [1, 2] show that it is possible to obtain good quality topic signatures for word senses automatically. [4] show that topic signatures for word senses can be used for extending WordNet's taxonomy, and [3] show that they are effective for clustering WordNet word senses.

1st. sense: **church, Christian_church, Christianity** *"a group of Christians; any group professing Christian doctrine or belief;"*

size 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) order(133.23) sect(127.85) old(86.11) greek(68.65) century(61.99) history(50.36) pentecostal(50.18) england(44.77) saint(40.23) america(40.14) holy(35.98) pope(32.87) priest(29.76) russian(29.75) culture(28.43) christianity(27.87) religious(27.10) reformation(25.39) ukrainian(23.20) mary(22.86) belong(21.83) bishop(21.57) anglican(18.19) rite(18.16) teaching(16.50) christian(15.57) diocese(15.44)

2nd. sense: **church, church_building** *"a place for public (especially Christian) worship;"*

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) welcome(29.23) new(28.94) prayer(24.48) temple(24.40) design(24.25) brick(24.24) erect(23.85) door(20.07) heaven(19.72) plan(18.26) call(17.99) renovation(17.78) mile(17.63) gate(17.09) architect(16.86) conservative(16.46) situate(16.46) site(16.37) demolition(16.16) quaker(15.99) fort(14.59) arson(12.93) sultan(12.93) community(12.88) hill(12.62)

3rd. sense: **church_service, church** *"a service conducted in a church;"*

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) cremation(64.73) saturday(60.46) thursday(60.46) june(57.78) tuesday(56.08) crematorium(55.53) weekly(53.36) procession(50.53) burial(48.60) december(48.46) ceremony(46.47) september(46.10) interment(42.31) lead(38.79) family(34.19) deceased(31.73) visitation(31.44)

Fig. 1. Fragment of the topic signatures for the three senses of church built with the monosemous relatives method to extract examples from the Web. The values in parenthesis correspond to χ^2 values. Only the top scoring terms are shown.

In this paper we compare similarity measures for WordNet concepts based on topic signatures with measures based on the hierarchy WordNet (see [5, 6] for recent references). The advantage of topic signatures over the similarity measures based on the hiearchy of WordNet is that they can be applied to unknown concepts, and thus allow for classifying new concepts. We also compare the impact of different ways of acquiring and modeling topic signatures.

The paper is structured as follows. Section 2 presents the method to construct topic signatures, alongside some parameters for the construction of them. In section 3 we review different methods to compute the similarity between topic signatures. Section 4 presents the experimental setting and Section 5 presents the results. Finally, Section 6 presents the conclusions and future work.

2 Construction of topic signatures

Two main alternatives for the construction of topic signatures have been presented in [1, 2] and [4], which will be presented briefly in this section. Please refer to those papers for further details. The first step consists on acquiring examples for the target word senses. The idea is to use the information in WordNet in order to build appropriate queries, which are used to search in the Internet those texts related to the given word sense. The second step organizes the examples thus retrieved in document collections, one collection for each word sense. In the third step, we extract the words in each of the collections and their frequencies, and compare them with the data in the other collections. Finally, The words that have a distinctive frequency for one of the collections are collected in a list, which constitutes the topic signature for each word sense. The steps are further explained below.

2.1 Acquiring the examples and building the document collections

In order to retrieve documents that are associated to a word sense, [1, 2, 4] present different strategies to build the queries. Some of the methods that were used have problems to scale-up, as they require certain amount of hand correction, so we propose to use two simple methods to build queries:

1. use all relatives (synonyms, hyponyms, children, siblings) of the target word sense
2. use only those relatives of the target word sense that are monosemous

One can argue that the first method, due to the polysemy of the relatives, can gather examples of relatives which are not really related to the target word sense. In principle, the second method avoids this problem and should provide better examples.

In the current implementation 1) was performed retrieving up to 100 documents from Altavista, and extracting from them the sentences which contain any of the synset words; and 2) was performed retrieving up to 1000 sentences for each monosemous relative from Google snippets.

2.2 Representing context

In order to model the retrieved examples we can treat the context as a bag of words, that is, all the words in the context are used in flat vector. In this case we build a vector of V dimensions (where V is the size of the vocabulary), where the words occurring in the contexts are the keys and their frequency the values. All the words are first lemmatized.

2.3 Weighting the words in context

Frequencies are not good indicators of relevancy, so different functions can be used in order to measure the relevance of each term appearing in the vector corresponding to one sense in contrast to the others. That is, terms occurring frequently with one sense, but not with others, are assigned large weights for the associated word sense, and low values for the rest of word senses. Terms occurring evenly among all word senses are also assigned low weights for all the word senses. We have currently implemented five measures: two versions of tf.idf[1] , χ^2, mutual information and t-score.

The topic signatures are vectors where the words have weights corresponding to the relevancy functions thus computed.

2.4 Filtering

In [2] it is shown that weighting functions can assign very high weights to rare terms appearing in the context of one of the word senses by chance. This effect can be reduced in the following way: we collect contexts of occurrences for the target *word* from a large corpus, and select the words that are highly related to the word. This list of words related to the target word is used in order to filter all topic signatures corresponding to the target word, that is, context terms which are not relevant for the target word are deleted from the topic signature. We have tested both filtered and unfiltered settings.

3 Similarity measures

Once we have constructed the topic signatures, it is possible to calculate the similarity between word senses using their topic signatures. If every word which can appear in a topic signature is considered a dimension in a Euclidean space, the similarity between two topic signatures can be calculated using the cosine of the angle of the vectors, or the Euclidean distance between them[2].

In order to evaluate the quality of our similarity measures we have taken two similarity metrics based on the WordNet hierarchy and used them as gold standards.

- Resnik's distance metric based on the Information Content of the synset [10].
- The inverse of the minimal number of hypernymy links between the two synsets in the WordNet hierarchy, also called *Conceptual Distance.*

[1] (a) $\frac{tf_t}{max_t tf_t} \times \log \frac{N}{df_t}$ (b) $(0.5 + \frac{0.5 \times tf_t}{max_t tf_t}) \log \frac{N}{df_t}$

[2] We have calculated the Euclidean distance of the unnormalized vectors because, in our experiments, a normalization produced that all the distances between the signatures became very similar and there was not much difference between the different weight functions used.

Besides, we have also taken the manually defined coarse-grained senses used in the Word Sense Disambiguation exercise Senseval-2. In order to define a similarity matrix based on this resource, we have considered two synsets similar if they are in the same coarse-grained sense (similarity 1), and dissimilar otherwise (similarity 0).

4 Experimental setting

Each experiment has been performed with a different choice of parameters in the construction of topic signatures:

1. Building the collections (monosemous vs. all relatives)
2. Weight function (χ^2, tf·idf, MI or t-score)
3. With or without filtering
4. Similarity metric between the topic signatures: cosine or Euclidean.

The evaluation has been done with sixteen nouns from the Senseval 2 exercise that were also used in [3] (WordNet version 1.7).

The correlation between our proposed similarity measures and the three gold standard similarity measures was used as a quality measure. The correlation was computed in the following way. First, for every noun, a symmetric similarity matrix is calculated containing the gold standard similarity between each pair of senses, and another matrix is calculated using the topic signatures. The correlation between the two matrices has been calculated transforming the matrices into vectors (after removing the diagonal and the values which are duplicated because of its symmetry) and using the cosine between the vectors. A measure of 1 will give us perfect similarity, in contrast to a measure of 0.

5 Results

Table 1 shows the results for the sixteen words separately and overall, using monosemous relatives for collecting the documents, the cosine similarity between the topic signatures, and the link distance in WordNet as gold standard. In the case of the word *dike*, the similarities are always 0 or 1. This is due to the fact that it only has two senses in WordNet. Therefore, there is only one similarity value between the two senses, and the cosine similarity between using a theoretical metric and using the topic signatures is 1 when both values are non-zero, and 0 when one of the values happens to be 0 (as it is the case when both topic signatures have no word in common). The best results are obtained for unfiltered topic signatures where MI is used as the weighting function.

Tables 2 and 3 list the results obtained for each possible configuration. The results show that it is possible to approximate very accurately the similarity metric based on link distance, as it is possible to attain a similarity of 0.88 with monosemous relatives and the MI or the t-score weight functions. The similarity between Resnik's function and the signatures was somewhat lower, with a cosine

Table 1. Similarity values, using the monosemous relatives queries, the cosine similarity for comparing the signatures, and correlation to the link distance in WordNet as gold standard.

Weight:	**Chi2**		**Tf·idf$_1$**		**Tf·idf$_2$**		**MI**		**t-score**	
Filtering:	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
art	0.85	0.67	0.71	0.83	0.82	0.82	**0.99**	**0.99**	0.58	0.59
authority	0.18	0.17	0.72	0.82	0.73	0.74	0.83	**0.85**	0.53	0.42
bar	0.3	0.31	0.53	0.45	0.64	0.66	**0.79**	0.74	0.15	0.05
bum	0.79	0.77	0.92	0.74	0.85	0.41	**1**	0.99	0.72	0.75
chair	0.48	0.41	0.78	0.71	0.78	0.8	**0.98**	0.91	0.61	0.67
channel	0.44	0.44	0.62	0.64	0.83	0.87	**0.92**	**0.92**	0.41	0.66
child	0.28	0.26	0.84	0.85	0.79	0.81	0.62	0.64	0.87	**0.89**
church	0.7	0.7	0.97	0.89	0.9	0.88	0.98	**1**	**1**	**1**
circuit	0.6	0.53	0.62	0.61	0.79	0.81	**0.97**	0.96	0.58	0.42
day	0.5	0.54	0.5	0.52	0.7	0.77	0.91	**0.92**	-0.02	0.04
dike	0	0	0	0	0	0	1	1	1	1
facility	0.41	0.54	0.68	0.64	0.7	0.72	**0.91**	0.78	0.47	0.39
fatigue	0.82	0.81	**0.92**	0.43	0.52	0.48	0.68	0.56	0.43	0.37
feeling	0.26	0.27	0.31	0.35	0.82	**0.87**	0.83	**0.87**	0.62	0.68
grip	0.29	0.24	0.8	0.66	0.7	0.76	0.86	**0.89**	0.43	0.45
hearth	0.8	0.68	0.75	0.92	0.71	0.8	0.96	**1**	0.89	0.89
MEAN	0.47	0.45	0.65	0.6	0.69	0.72	**0.88**	0.87	0.61	0.62

similarity of 0.65 (again with the MI weight function, but with the all relatives signature). Finally, it was more difficult to approximate the similarity based on the coarse grained senses, as it does not provide similarity values in $\Re$ but binary values. Nonetheless, it was possible to obtain a cosine similarity of 0.47 with a tf·idf function.

Regarding the parameters of topic signature construction, the monosemous relative method obtains the best correlation when compared to the link distance gold standard. As this method uses a larger amount of examples than the all relatives method, we cannot be conclusive on this. Previous experiments [1] already showed that short contexts of larger amount of examples were prefereable rather that larger context windows and fewer examples. On the same gold standard, MI and t-score attain much better correlation scores that the rest of weighting functions. Filtering the topic signature does not improve the results, and both Euclidean distance and the cosine yield the same scores.

6 Conclusion and future work

The experiments show that it is possible to approximate accurately the link distance between synsets (a semantic distance based on the internal structure of WordNet) with topic signatures. However, Resnik's metric [10] has not been as easily captured by the topic signatures, so more work is needed to be able to

Table 2. Results for the signatures obtained with the monosemous relatives procedure, given as correlation measures against each of the three gold standards

	Weight	*Chi2*		*Tf·idf*$_1$		*Tf·idf*$_2$		*MI*		*t-score*	
	Filtering	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
Gold std.	**Metric**										
Coarse-grained	*Euclidean*	0.14	0.12	0.25	0.23	0.19	0.08	0.3	**0.33**	**0.33**	0.29
senses	*cosine*	0.22	0.21	0.38	**0.47**	0.34	0.37	0.37	0.39	0.17	0.2
Resnik	*Euclidean*	0.31	0.28	0.35	0.44	0.28	0.39	**0.56**	**0.56**	0.55	0.51
	cosine	0.39	0.38	0.35	0.26	0.35	0.37	**0.52**	0.49	0.31	0.35
links	*Euclidean*	0.63	0.61	0.63	0.7	0.48	0.51	0.81	**0.87**	**0.87**	0.8
	cosine	0.47	0.45	0.65	0.6	0.69	0.72	**0.88**	0.87	0.61	0.62

Table 3. Results for the signatures obtained with the all relatives procedure, given as correlation measures against each of the three gold standards

	Weight	*Chi2*		*Tf·idf*$_1$		*Tf·idf*$_2$		*MI*		*t-score*	
	Filtering	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
Gold std.	**Metric**										
Coarse-grained	*Euclidean*	0.17	0.16	0.18	0.18	0.18	0.18	0.33	0.33	**0.35**	0.32
senses	*cosine*	0.33	0.3	0.33	**0.39**	0.34	**0.39**	0.32	0.34	0.03	0.04
Resnik	*Euclidean*	0.38	0.37	0.3	0.3	0.3	0.3	0.48	**0.49**	**0.49**	0.47
	cosine	0.44	0.28	0.47	0.46	0.51	0.48	**0.65**	0.61	0.42	0.3
links	*Euclidean*	0.65	0.65	0.57	0.57	0.57	0.57	0.82	0.84	**0.85**	0.82
	cosine	0.49	0.43	0.62	0.62	0.68	0.69	0.81	**0.84**	0.44	0.43

approximate it with distributional procedures. The main source of the difference is that Resnik's metric gives a similarity of 0 to two synsets if they are located in different sub-taxonomies (with a different root node), such as *church* as a group, an entity or an act. On the other hand, there will probably be some similarity between the topic signatures of two such synsets. Finally, the gold standard metric based on the coarse-grained senses was the one that produced the lowest results. This is in clear contradiction with our word sense clustering experiments [3], where the clusters constructed using topic signatures replicated very well the coarse-grained senses. We think that the correlation metric is not a very appropriate evaluation method in this case, as any similarity metric will yield low correlation when compared to a boolean similarity metric.

Regarding the parameters for the construction of topic signatures, using monosemous relatives allows for better results. Contrary to our intuitions, filtering did not improve performance, and both Euclidean distance and the cosine yielded similar results. It has been a surprise that Mutual Information and t-score have provided much better results than other metrics, such as χ^2 and tf·idf, which have been used extensively for generating topic signatures in the past.

The next step in these experiments consists in ascertaining whether the settings for which the similarity has been better are also more useful when applied to the classification of new concepts, word sense disambiguation or word sense clustering.

Some other ideas for future work are:

- Compare to other similarity measured using WordNet [5]
- Repeat the experiment with other kinds of topic signatures, such as modeling the syntactic dependences between the synset considered and the context words [4].
- Explore further parameters in topic signature construction.

References

1. E. Agirre, O. Ansa, E. Hovy, and D. Martinez. Enriching very large ontologies using the WWW. In *Ontology Learning Workshop, ECAI*, Berlin, Germany, 2000.
2. E. Agirre, O. Ansa, D. Martínez, and E. Hovy. Enriching wordnet concepts with topic signatures. In *Proceedings of the SIGLEX workshop on WordNet and Other lexical Resources: Applications, Extensions and Customizations, in conjunction with NAACL*, Pittsburg, 2001.
3. E. Agirre and O. Lopez de Lacalle Lekuona. Clustering wordnet word senses. In *Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP'03)*, Bulgary, 2003.
4. E. Alfonseca and S. Manandhar. Extending a lexical ontology by a combination of distributional semantics signatures. In *Proceedings of EKAW'02*, Siguenza, Spain, 2002. Also published in *Knowledge Engineering and Knowledge Management*. Lecture Notes in Artificial Intelligence 2473. Springer Verlag.
5. P. Banerjee and T. Pedersen. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, February 2003.
6. A. Budanitsky and G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, June 2001.
7. C. Fellbaum. *Wordnet: An Electronic Lexical Database*. Cambridge: MIT Press, 1998.
8. C.-Y. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In *Proc. of the COLING Conference*, Strasbourg, France, August 2000.
9. A. Maedche and S. Staab. Ontology learning. In S. Staab and R. Studer, editors, *Handbook of Ontologies in Information Systems*. Springer Verlag, Forthcoming.
10. P. S. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.