Comparing methods for automatic acquisition of Topic Signatures

Montse Cuadros, Lluis Padro
TALP Research Center
Universitat Politecnica de Catalunya
C/Jordi Girona, Omega S107
08034 Barcelona
{cuadros, padro}@lsi.upc.es

German Rigau
IXA Group
Euskal Herriko Unibertsitatea
P.Manuel Irdiazabal, 1
20018 Donostia
rigau@si.ehu.edu

Abstract

The main goal of this work is to compare two methods for building Topic Signatures, which are vectors of weighted words acquired from large corpora. We used two different software tools, ExRetriever and Infomap, for acquiring Topic Signatures from corpus. Using these tools, we retrieve sense examples from large text collections. Both systems construct a query for each word sense using WordNet. The quality of the acquired Topic Signatures is indirectly evaluated on the Word Sense Disambiguation English Lexical Task of Senseval-2.

keywords: Topic Signatures, acquisition, Latent Semantic Indexing, Word Sense Disambiguation, Multilingual Central Repository, WordNet.

1 Introduction

Topic Signatures (TS) are word vectors related to a particular topic. Topic Signatures are built by retrieving context words of a target word from large text collections. They have been used in a variety of ways, such as in Summarization Tasks (Lin & Hovy 00), ontology population (Alfonseca *et al.* 04) or word sense disambiguation (Agirre *et al.* 00), (Agirre *et al.* 01). In fact, there is now available Topic Signatures for all WordNet (Fellbaum 98) nominal senses (Agirre & de laCalle 04).

This work presents a comparison of two different techniques for building Topic Signatures.

The first technique retrieve contexts using queries which consist of a set of literal words. Although these systems have been improved with several enhancements such as term weighting, authority linking, and ad-hoc heuristics to improve their performance, these lexical matching methods can be inaccurate because the queries are based on words instead of concepts. However, there are many ways to characterize a given concept.

The second technique uses Latent Semantic Indexing (LSI). LSI tries to overcome the problems of lexical matching by using statistically derived conceptual indexes instead of literal words for retrieval. This technique assumes that there is some underlying latent semantic structure in the data.

A Topic Signature, for our purposes, is a weighted vector of words related to a particular word sense. We tried two different systems for build Topic Signatures. The first one, ExRetriever (Cuadros *et al.* 04), is based on the first technique described above, and the second one, Infomap (Dorow & Widdows 03), is based on the second technique.

Our main goal with this study, as mentioned before, is to compare the performances of both methods for automatic TS acquisition. In order to perform this comparison, we evaluated the TS acquired by both systems in a specific task, the English-Lexical Sample task of Senseval-2.

For building the Topic Signatures for all the word senses of the Lexical Sample task of Senseval-2 we used BNC (British National Corpus).

This paper is organised as follows: In section 2, we explain in detail the software tools we use for the task, providing a brief explanation of Latent Semantic Indexing (LSI). In section 3, we explain the steps followed to construct the Topic Signatures and in section 4, the results of the indirect evaluation we carried out. Finally, in section 5 and 6, some concluding remarks and future work are provided.

2 Tools

2.1 ExRetriever

ExRetriever is a flexible tool to perform sense queries on large corpora (Cuadros *et al.* 04). ExRetriever characterises automatically each synset of a word as a query (mainly using: synonyms, hyponyms and the words of the definitions); and then, uses these queries to obtain sense examples (sentences) automatically from a large text collection. The current implementation of

ExRetriever accesses directly the content of the Multilingual Central Repository (MCR) (Atserias *et al.* 04) of the MEANING project which includes several WordNet versions. The system uses also SWISH-E[1] to index large collections of text such as SemCor (Miller *et al.* 93) or BNC. SWISH-E is a fast, powerful, flexible, free, and easy to use system for indexing collections of Web pages or other files. ExRetriever has been designed to be easily ported to other lexical knowledge bases and corpora, including the possibility to query search engines such as Google.

2.2 Infomap

The Infomap NLP Software package[2] uses a variant of Latent Semantic Indexing (LSI) on free-text corpora to learn vectors representing the meanings of words in a reduced vector-space known as Word-Space (Dorow & Widdows 03).

The Infomap software performs two basic functions: building models by learning them from a free-text corpus using certain learning parameters specified by the user, and searching an existing model to find the words or documents that best match a query according to that model.

2.2.1 Latent Semantic Indexing

Latent Semantic Indexing (LSI) allows to extract and represent the contextual meaning of words by statistical computations applied to a large corpus of text (Schtze 98). The underlying idea is that when reducing the dimensionality of the original word-space, similar words are projected closer to each other in the reduced space while dissimilar words are projected to distant locations. The reduced space is obtained using linear algebra methods, in particular, the Singular Value Decomposition (SVD). Part of the motivation for using SVD for word vectors is the success of LSI in information retrieval.

Latent Semantic Indexing maps the contextual relationships between words in terms of common usage across a collection of documents. LSI enables to understand how words relate to each other through the creation of a similarity measure, which reveals whether a given word or document is similarly used compared with another word or document.

[1] `http://swish-e.org`

[2] `http://infomap-nlp.sourceforge.net/`

3 Strategies for acquiring Topic Signatures

In order to evaluate the performance of both approaches, we designed a preliminary set of strategies for acquiring the Topic Signatures from BNC.

3.1 Acquisition Process

The acquisition process consist of the following steps:

1. Devise a particular strategy for query construction and apply the query construction schema to all the senses of a word.
2. Perform the sense queries on the BNC.
3. Collect the sense corpus.
4. Obtain a Topic Signature for each sense.

3.2 Query construction strategies

We have designed a few preliminary set of query construction strategies based on synonymy, hyponymy and hypernymy relationship of WordNet inspired by the work of (Leacock *et al.* 98).

- A) Monosemous strategy : (OR monosemous-words) the union set of all the synonym, hyponym and hyperonym monosemous words of a WordNet sense.
- B) Polysemous strategy : (OR polysemous-words) the union set of all the synonym, hyponym and hyperonym polisemous words of a WordNet sense.
- C) Monosemous and Polysemous strategy : (word AND (OR polysemous-words)) OR* (OR monosemous-words) the union set of all synonym, hyponym and hyperonym monosemous and polisemous words of a WordNet sense in such a way. OR* stands for a particular OR boolean function to express that there is at least one monosemous word or the word and one polysemous word.

We remove those words (monosemous or polysemous) appearing in more than one sense query, trying to construct the sense queries in such a way, that there is no overlapping words in different sense queries of the same word.

3.3 Construction of the Topic Signatures using ExRetriever

These queries have been applied to locate particular sentences of the BNC using ExRetriever. In that way, we are able to retrieve a set of examples for each word sense. In all cases, we remove all stop words from the corpus. Afterwards, we calculate the Mutual Information for each word in the sense corpus with respect to their synset using the formula (1).

$$MI(w,s) = log\frac{P(w \wedge s)}{P(w)P(s)} \quad (1)$$

Given a word w and their word sense s, $P(w \wedge s)$ represents the probability of appearing w in the s sense. $P(w)$ is the probability of occurring w in the BNC corpus, and $P(s)$ is the probability of a document (sentence) to belong to the s sense.

As an example, we will show the full process of obtaining a Topic Signature.

For example, a query of type C for the word *church#n* is constructed using ExRetriever as follows:

As WordNet 1.7 *church#n* has three senses, ExRetriever builds three different queries:

- sense 1: ((church and (christianity or protestant or religion)) or christian_church or catholic_church or coptic_church)
- sense 2: ((church and (abbey or basilica or cathedral)) or church_building or kirk or place_of_worship or house_of_prayer or house_of_god)
- sense 3: ((church and (service)) or church_service or religious_service or divine_service)

Once we construct each sense query, we use ExRetriever to gather all matching sentence examples from the BNC corpus. Afterwards, we calculate the Mutual Information of all the words appearing in the corpora obtained.

We have calculated the Topic Signatures for query A and C, in an improved method based on not taking account the case of the words and looking for the appearance of the exact compound-words in the gathered examples.

After this process, we obtain per each word sense, a word vector with weights (Topic Signatures). Table 1 presents some resulting words for sense 3 of *church#n* using the strategy A).

witness	2.229616	context	2.411937
burial	2.517298	husband	2.517298
participants	2.517298	sermon	2.517298
service	2.715123	adapted	2.715123
adults	2.715123	afternoon	2.922763
agenda	2.922763	arranged	2.922763
attracted	2.922763	audible	2.922763
augment	2.922763	award	2.922763

Table 1: Example of a Topic Signature obtained with ExRetriever

3.4 Construction of the Topic Signatures using Infomap

Infomap only allows AND and ANDNOT operator and does not consider the OR operator. For this reason, the queries have been modified slightly. We use the same words that we used when querying with ExRetriever but we remove all the operators (by default Infomap uses the AND operator).

After building a model with the corpus, the *associate* command of Infomap can return both a list of the words or the documents best matching the query, in descending order of relevance. Using this option provided by Infomap, once we have the queries, we obtain the list of weighted words that in this experiment we consider the Topic Signature of the query. Table 2 presents the resulting words for sense 3 of *church#n* using the strategy C) with higher relevance.

service	0.776187	anglican	0.651298
church	0.776186	services	0.651127
clergy	0.718070	tower	0.651071
hymns	0.695500	st	0.650787
peterś	0.695215	congregational	0.648595
episcopal	0.689341	congregation	0.647037
presbyterian	0.685548	priest	0.644656
cathedral	0.685220	memorial	0.644652
churches	0.683878	charters	0.642540
royal	0.673297	worship	0.637472
parish	0.671534	bishop	0.634107
pastoral	0.670789	volunteer	0.629541
maryś	0.666601	...	

Table 2: Example of a Topic Signature obtained with Infomap

4 Indirect evaluation on Word Sense Disambiguation

In order to measure the quality of the acquired TS by these two different approaches, we performed an indirect evaluation by using the acquired Topic Signatures (TS) for a Word Sense Disambiguation (WSD) task. In particular, the Senseval-2

English Lexical Sample task. We used this evaluation framework instead of the the one provided by Senseval-3 because in this case, the verbal part was not directly annotated using WordNet senses.

The TS are applied to all the examples of the test set of the Senseval-2 using a simple word overlapping (or weighting) counting. That is, the program calculates the total number of overlapping words between the Topic Signature and the test example. The sense having higher counting (or weighting) is selected for that particular test example. In table 3, we can see an example of the evaluation test corresponding to sense 3 of *church#n*. As we can see, in bold there are some words that appear in the Topic Signatures for sense 3 obtained using Infomap.

In table 4 appears a summary of the results of this indirect evaluation. This table presents the results for each type of query construction strategy (either A, B or C), each system (either Infomap or ExRetriever), and with several levels of sense granularity (either fine or coarse). In this table, P stands for Precision, R for Recall and F1 for F1 measure.

The best figures are obtained by using the Infomap method with occurrences, which is not surprising due to the LSI effect (39.1 precision and recall for fine grained granularity). In table 4, we present the official results of the Senseval-2 of those systems declared to be unsupervised. When comparing with those systems, Infomap would score second while ExRetriever fourth getting as a reference the recall in fine-grained. Looking at literature, (Agirre & Martinez 04), UNED-LS-U unsupervised method is considered semi-supervised. This approach, uses some heuristics rely on the bias information available in Semcor. The distribution of senses is used to discard low-frequency senses.

In table 4, we present the results of the queries for each system based on POS, and we can see that the best query for each POS always rely on A, the only difference is that sometimes the best result uses the occurrence or the weight measure method. We have put the results of the improved methods for ExRetriever. If we had used the best method for each part of speech, we had improved our results achieving a precision of 31.5, a recall of 29.7 and a f1 of 30.57 which would imply to be one position over in the 4 results for ExRetriever. Otherwise, Infomap would improve not very significantly, we would get a precision and recall of 39.3, that would mean that we would be in the same position.

As expected, regarding the query construction strategy, in general it seems that strategy A (Monosemous strategy), is better than C (Monosemous and Polysemous strategy) and B (Polysemous strategy), which is the one with the lowest results. We also obtain similar figures with respect occurrences vs. weights methods: using Infomap we obtain slightly better figures for occurrences while when using ExRetriever the best results appear for weights.

5 Conclusions

We presented some experiments using two software tools to compare the automatic acquisition of Topic Signatures for word senses. Our Evaluation Framework has been the English Lexical Sample task of Senseval-2. We have focus on the Senseval-2 task because it uses the synsets of WordNet 1.7 for each part of speech, and then is more reliable to our experiments because our queries are build with WordNet 1.7.

We can observe that using Infomap, the tool developed to work with vector models acquired from Corpus, we obtain promising results.

In order to improve the ExRetriever results we plan to filter out those words that seem to be very common in all senses, for example, Named Entities, Multi Words Expressions, etc. or keeping those words that have a common domain or any other semantic relation in common.

Infomap vectors seem to be more accurate for obtaining good context words of an specific word sense. Furthermore, it seems that the results could improve largely varying different system parameters such as dimensionalty of the model, size of the Topic Signatures, etc.

We also plan to tune separately each part-of-speech.

6 Acknowledgements

This work is supported by the European Commission (MEANING IST-2001-34460). Our research group, TALP Research Center, is recognized as a Quality Research Group (2001 SGR 00254) by DURSI, the Research Department of the Catalan Government.

In developing measuring tools for the local **church** we are concerned with quality control as much as quantity performance, to use commercial language. Responsible leaders want to know how people are growing in their understanding of the Christian faith, whether relationships are deepening and extending throughout the **church-fellowship**, and to what extent the Christian presence is evident in the community outside. Such information cannot be gathered with such precision as numerical data, but it is essential that each area be investigated to ensure that there is a balance between **worship**, fellowship, learning, evangelism and **service**. Healthy organic growth is proportionate, with each area and function developing in relation to the other. Quality of <*head*> **church** <*head*> life can be measured in the following three ways

Table 3: Test example for noun church

		fine			**coarse**		
Method	Query	P	R	F1	P	R	F1
Infomap	A	**39.1**	**39.1**	**39.1**	**51.0**	**51.0**	**51.0**
occurrences	B	37.8	33.2	35.3	50.0	43.8	46.7
	C	37.8	33.2	33.2	50.0	43.8	46.7
Infomap	A	**39.1**	**39.1**	**39.1**	**50.7**	**50.7**	**50.7**
weights	B	38.4	32.8	35.4	49.9	42.7	46.02
	C	38.4	32.8	35.38	49.9	42.7	46.02
ExRetriever	A	**28.5**	**27.1**	**27.8**	**42.3**	**40.3**	**41.3**
occurrences	B	24.1	17.2	20.0	35.4	25.3	29.5
	C	21.7	21.3	21.5	36.6	36.0	36.3
ExRetriever	A	**28.9**	**27.2**	**28.02**	**41.9**	**39.3**	**40.6**
weight	B	22.6	15.9	18.67	33.0	23.2	27.3
	C	25.1	24.6	24.85	36.9	36.1	36.5

Table 4: Overall results of the systems using Senseval-2 with respect fine-grained and coarse-grained senses

Method	Query	Noun	Verb	Adj
Infomap	A	40.1	**32.2**	**53.3**
occurrences	B	34.26	29.47	51.29
	C	34.26	29.47	51.29
Infomap	A	**40.6**	31.7	53
weights	B	34.93	29.19	50.77
	C	34.93	29.19	50.77
ExRetriever	A	27.8	**28**	**27.03**
occurrences	C	25.3	17.1	22.79
ExRetriever	A	**34.6**	23.25	23.64
weights	C	32.45	18.2	23.39

Table 5: F1 related to each POS

References

(Agirre & de laCalle 04) E. Agirre and O. Lopez de la Calle. Publicity available topic signatures for all wordnet nominal senses. In *LREC'04*, pages 97–104, 2004.

(Agirre & Martinez 04) E. Agirre and D. Martinez. Unsupervised wsd based on automatically retrieved examples: The importance of bias. In *Proceedings of the EMNLP*, Barcelona, 2004.

(Agirre *et al.* 00) E. Agirre, O. Ansa, D. Martinez, and E. Hovy. Enriching very large ontologies with topic signatures. In *Proceedings of ECAI'00 workshop on Ontology Learning*, Berlin, Germany, 2000.

(Agirre *et al.* 01) E. Agirre, O. Ansa, D. Martez, and E. Hovy. Enriching wordnet concepts with topic signatures. In *Proceedings of the NAACL workshop on WordNet and Other lexical Resources: Applications, Extensions and Customizations*, Pittsburg, 2001.

(Alfonseca *et al.* 04) E. Alfonseca, E. Agirre, and O. Lopez de Lacalle. Approximating hierachy-based similarity for wordnet nominal synsets using topic signatures. In *Proceedings of the Second International Global WordNet Conference (GWC'04). Panel on figurative language*, Brno, Czech Republic, January 2004. ISBN 80-210-3302-9.

	fine			coarse		
Method	P	R	F1	P	R	F1
UNED - LS-U	40.2	40.1	40.15	51.8	51.7	51.75
Infomap - A Occ	**39.1**	**39.1**	**39.1**	**51.0**	**51.0**	**51.0**
ITRI - WASPS-Workbench	58.1	31.9	41.19	66.1	36.3	46.86
CL Research - DIMAP	29.3	29.3	29.3	36.7	36.7	36.7
ExRetriever - A weight	**28.9**	**27.2**	**28.02**	**41.9**	**39.3**	**40.56**
IIT 2 (R)	24.7	24.4	24.55	34.6	34.1	34.35
IIT 1 (R)	24.3	23.9	24.1	34.1	33.6	33.85
IIT 2	23.3	23.2	23.25	32.3	32.2	32.25
IIT 1	22	22	22	32.1	32	32.05

Table 6: Senseval-2 systems results for fine-grained and coarse-grained senses, in wining order

(Atserias *et al.* 04) Jordi Atserias, Luís Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen. The meaning multilingual central repository. In *Proceedings of the Second International Global WordNet Conference (GWC'04)*, Brno, Czech Republic, January 2004. ISBN 80-210-3302-9.

(Cuadros *et al.* 04) M. Cuadros, M. Castillo, G. Rigau, and J. Atserias. Automatic Acquisition of Sense Examples using ExRetriever. In *Iberamia'04*, pages 97–104, 2004.

(Dorow & Widdows 03) B. Dorow and D. Widdows. Discovering corpus-specific word senses. In *EACL*, Budapest, 2003.

(Fellbaum 98) C. Fellbaum, editor. *WordNet. An Electronic Lexical Database*. The MIT Press, 1998.

(Leacock *et al.* 98) C. Leacock, M. Chodorow, and G. Miller. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1):147–166, 1998.

(Lin & Hovy 00) C. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of 18th International Conference of Computational Linguistics, COLING'00*, 2000. Strasbourg, France.

(Miller *et al.* 93) G. Miller, C. Leacock, R. Tengi, and R. Bunker. A Semantic Concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, 1993.

(Schtze 98) H. Schtze. Automatic word sense discrimination. In *Computational Linguistics*, 1998.