Automatic Acquisition of Domain Specific Lexicons

The IST Programme
Shared-Cost RTD
MEANING
Developing Multilingual Web-scale Language Technologies
Contract Number: IST-2001-34460

Workpackage: WP5
Deliverable: WP5.12 - 5.C
Version: Working Paper

Authors: Alfio Gliozzo, Carlo Strapparava, Ernesto D'Avanzo, Bernardo Magnini - ITC-irst

Date: 25 January 2005

Contents

Abstract

In this paper we present the results of three years of experiments about automatic acquisition of domain specific terminology from corpora. We present an analysis of the potentiality and limitations of the Term Categorization approach to lexical acquisition, and we propose a novel methodology to approach the task, consisting on applying Latent Semantic Kernels to estimate term similarity. We find out that domain specific monosemous terms behave similarly to domain specific lexical items, so we used them to train and evaluate our Term Categorization system. Results show that the proposed technique is effective, achieving an accuracy of about 43%. We also reported an error analysis showing that most of the misclassification errors are related the the fuzzy nature of domain distinctions. In particular we identified a set of "families" in the WordNet Domains categories that makes difficult the classification task. Categorizing monosemous terms according to domain labels allows us to automatically assigning domain labels to a subset of the WordNet synsets, allowing to perform a bootstrap procedure to assign a of domain label to every synset in WordNet.

1 Introduction

In this paper we present the results of three years of experiments about automatic acquisition of domain specific terminology from corpora. All the experiments have been performed in the context of the Meaning project. The original goal of our research line was to study the possibility of automatically acquiring domain specific lexicon from corpora, adopting a Term Categorization methodology [Avancini *et al.*, 2003; Lavelli *et al.*, 2002]. In particular, our original plan concentrated on acquiring lists of domain specific Named Entities from corpora.

Before starting the acquisition methodology, we did an analysis of the potentiality and limitations of the Term Categorization approach to lexical acquisition, and we found some very crucial limitation, that makes it inadequate for a large scale lexical acquisition process. In particular the TC algorithm proposed in [Avancini *et al.*, 2003] is affected by a very low recall, expecially for terms having a low nomber of occurrences in the corpus. Most of domain specific terms are not frequent in the corpus, and many of them are very sparse. Then we performed a deeper analysis of the task, to identify and to solve these problems before starting the acquisition process.

As a solution we propose the use of Latent Semantic Kernels to approach the TC task. The main advantages of this approach is that the sparseness and the dimensionality problem (see section 3) are solved together by adopting a cluster based vector space model to represent terms (e.g. Latent Semantic Indexing (LSI) [Deerwester *et al.*, 1990]). Then we exploited a LSI representation for the feature space of terms, and we trained a Support Vector Machine (SVM) classifier on it.

Then we concentrated on analyzing the evaluation task. In the literature [Avancini *et al.*, 2003; Lavelli *et al.*, 2002] all the nouns in WordNet Domains have been adopted as a Gold Standard dataset for both training and testing the TC system. We found that this

approach is inadequate: more than one domain is often assigned to ambiguous words , and most of them are not domain specific terms.

On the other hand, domain specific terms are not likely to be neither ambiguous nor generic. Then the set of all the non generic monosemous terms contained in WordNet can be used to train and to evaluate a TC system with the expectation that the estimated performances will be reflected in the "real "acquisition task.

An automatic technique to acquire domain labels for synsets in WordNet can be used for several purposes. For example the lexical resource can be tuned for a particular domain by pruning the irrelevant senses for the domain. Another useful application of TC is ontology population. In this case list of terms not yet contained in the lexical resource can be acquired from a corpus. Then they can be automatically labeled by the TC algorithm, and included in the lexical resource in the correct "domain area". For example named entities in texts, such as *Kasparov*, can be related to synsets belonging to a specific domain (in this case Chess).

Experimental results show that applying Latent Semantic Kernels to the TC task is an effective strategy: the TC system achieves an accuracy of about 43% for all the monosemus terms in a corpus, with a very good improvement in recall, if compared with the results that can be found in the literature [Avancini *et al.*, 2003; Lavelli *et al.*, 2002]. We also reported an error analysis showing that most of the misclassification errors are related the the fuzzy nature of domain distinctions. In particular we identified a set of "families" in the WordNet Domains categories that makes difficult the classification task.

Due to limitation of time we were not able to evaluate the acquisition procedure in a "real" lexical acquisition task. Anyway, we improved substantially the understanding of the TC problem, and then its state of the art.

The paper is structured as follows. In section 2 we introduce the concept of semantic domains. Section 3 describes the use of Latent Semantic Kernels (LSK) for TC. Section 4 illustrates the resources we used for our experiments, while in section 5 we experimentally measured the viability of our approach to acquire domain information for monosemous words. Finally section 6 contains some final remarks.

2 Semantic Domains

This section introduces the notion of semantic domains from the computational linguistics perspective, suggesting that semantic domains provide a useful component for modeling conceptual structures.

Semantic domains are groups of strictly related concepts in the language, whose fundamental property is to frequently co-occur together in texts. Assigning semantic domains to concepts in a semantic network such as WordNet is then useful to define domain specific substructures of it, that can be profitably exploited as a basis for a further ontology learning process.

Domain information has been demonstrated to be very useful for lexical ambiguity resolution [Magnini *et al.*, 2002]. In particular knowing in advance the domain of the

context in which an ambiguous word is located allows to reduce sensibly its polysemy, making easier the overall disambiguation process.

An important property of the semantic domains is that they have a *dual* role in linguistic descriptions: semantic domains can be represented by both clusters of words and clusters of texts, exploiting a lexical coherence assumption [Gliozzo *et al.*, 2004]. Similarities among words can be estimated by considering their co-occurrences in texts, allowing to perform a large scale corpus based acquisition process for domain clusters of words and texts. Exploiting domain clusters is the preliminary step to acquire domain information for concepts. In the rest of this paper we will describe in details how domain clusters can be induced from corpora with a TC technique to associate domain labels to terms in a corpus.

3 Latent Semantic Kernels for Term Categorization

Term Categorization (TC) is the task of assigning domain labels to terms. In the literature [Avancini *et al.*, 2003; Lavelli *et al.*, 2002] this task has been approached as a supervised categorization problem: a preliminary set of terms have been labeled by adopting a predefined set of domains, and a supervised classifier has been trained on them. This approach allows us to automatically identify the semantic domains of new terms, achieving a reasonably good precision. The preliminary set of labeled terms has been extracted from WordNet Domains, as described in section 4.

In order to apply Machine Learning techniques to the TC task it is necessary to describe each term by a feature vector. Features are extracted by exploiting the information taken from a large scale corpus, containing some occurrence of the term. [Avancini *et al.*, 2003; Lavelli *et al.*, 2002] described each term by the vector of its frequencies in each document in the corpus. Thus each document is a dimension of a feature space. With this approach it is possible to achieve a good precision in classification (about 71% of precision and 8% of recall) for a subset of frequent terms (more than 10 documents per term). Surprisingly the authors reported a drop off in the performance when the number of documents in the corpus increases, in contrast to the intuition that augmenting the information about word usages makes TC easier.

The low recall, the high term frequency required and the impossibility of enlarging the corpus size make the TC approach infeasible for large scale acquisition of domain specific terminology. Most of domain specific terms are not frequent in the corpus, and many of them are very sparse. Then we performed a deeper analysis of the task, to identify and to solve these problems before starting the acquisition process.

An explanation for the drop off in accuracy when the dimensionality of the feature space increases is that increasing the corpus size introduces new dimensions in the vector space model representing the terms. The PAC theory [Kearns and Vazirani, 1994] predicts that the classification error increases with the dimensionality of the feature space. It is the case in which the size of the corpus used to perform TC is very large. In addition rare terms are sparse, so that unfrequent terms seldom co-occur in the same document, even

though they are semantically related.

The sparseness and the dimensionality problem can be solved together by adopting a cluster based vector space model to represent terms (e.g. Latent Semantic Indexing (LSI) [Deerwester *et al.*, 1990]). Then we exploited a LSI representation for the feature space of terms, and we trained a Support Vector Machine (SVM) classifier on it.

More formally let $T = \{t_1, t_2, \ldots, t_{|T|}\}$ be the corpus used for learning, let $V = \{w_1, w_2, \ldots, w_{|V|}\}$ be its vocabulary and let $f(w, t)$ be the frequency of the word w in the text t. The term by document matrix D has elements $d_{i,j} = f(w_i, t_j)$.

The LSI space is defined by performing a Singular Value Decomposition (SVD) on D[1]. SVD decomposes D into 3 matrixes $D \simeq V\Sigma_k T'$ where Σ_k is the diagonal matrix containing the highest k eigenvalues of D. The LSI representation for the term t_i is the vector $\vec{t_i}$ composed by the i^{th} row of the matrix $V\sqrt{\Sigma_k}$ [2]. The parameter k is the dimensionality of the LSI space and can be fixed in advance[3]. For our experiments we defined an LSI space by performing a SVD on the Term by Document matrix extracted from a subset of the BNC corpus (see section 4), fixing k to 100.

An LSI representation for terms allows us to better estimate term similarity by taking into account second order relations among terms and documents. In the LSI literature the cosine vector similarity has been used to estimate term similarity, as described by equation 1.

$$sim(t_i, t_j) = cos(\vec{t_i}, \vec{t_j}) \tag{1}$$

The LSI representation for TC present several advantages. First of all the use of an LSI space avoids the sparseness problems, by taking into account second order relations between terms while estimating term similarity. Second, the corpus size is not a limitation anymore, being the number of dimensions fixed in advance. Third, it is possible to represent new terms, not already seen in the corpus, in the same LSI space used for training by exploiting pseudo-term representation techniques [Berry *et al.*, 1995]. This last property is of great interest for TC because it allows to train a TC system once for all from a corpus in order to classify new terms contained in different corpora. In this case the new terms are represented on the LSI space defined in the training phase.

For our experiments we used LIBSVM[4], a multiclass SVM that adopts a pairwise classification strategy, and we exploited the Latent Semantic Kernel defined by equation 2.

$$K(t_i, t_j) = \frac{< \vec{t_i}, \vec{t_j} >}{\sqrt{< \vec{t_i}, \vec{t_i} >}\sqrt{< \vec{t_j}, \vec{t_j} >}} \tag{2}$$

[1] For our experiments we used SVDPACK, an SVD package optimized for sparse matrices freely downloadable from `http://www.netlib.org/svdpack/`.

[2] The term vectors are rescaled by their eigenvalues, according to the standard representation for term vectors described in [Berry *et al.*, 1995].

[3] It is not clear how to choose the right dimensionality of the LSI space. In general, values in the range $[100, 400]$ are reported in the literature.

[4] The libsvm package is freely available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

4 Experimental Settings

The lexical resources required to perform a TC task are a lexical database containing a preliminary set of terms labeled by domains and a large scale corpus. For the experiments reported in this paper we exploited WordNet Domains and the British National Corpus, described in the following paragraphs.

WordNet Domains. WordNet Domains [Magnini and Cavaglià, 2000] is an extension of WordNet [Fellbaum, 1998], in which each synset has been manually annotated with one or more *domain labels* (e.g. Sport, Medicine, etc.). About 164 domain labels were selected from a number of dictionaries and then structured in a taxonomy according to their position in the (much larger) Dewey Decimal Classification system [Comaroni *et al.*, 1989], which is commonly used for classifying books in the libraries.

Some WordNet synsets do not belong to a specific domain but rather correspond to general language and may appear in any context. Such senses are tagged in WordNet Domains with a Factotum label, which may be considered as a "placeholder" for all other domains.

We chose to use a subset (41) of the domain labels in WordNet Domains. For example, Sport is used instead of Volley or Basketball, which are subsumed by Sport. This subset was selected empirically to allow a sensible level of abstraction without losing much relevant information, overcoming data sparseness for less frequent domains.

From WordNet Domains it is possible to find out domain labels for terms, simply by collecting the set of domains labels for each synset in which the term appear. In [Lavelli *et al.*, 2002] WordNet Domains has been used as a Gold Standard for the TC task in order to demonstrate the viability of the TC process for lexical acquisition.

POS	w	MS	MD	$MS \cap MD$
N	20706	12703	8065	7496
V	4872	1678	555	471
A	8885	4885	1173	1109
All	34463	19266	9793	9076
$freq > 10$	23820	11079	5331	4835
$freq > 50$	10648	3100	1462	1244
$freq > 100$	6923	1568	701	583
$freq > 500$	2270	296	97	74

Table 1: Words in the BNC corpus

The British National Corpus. In order to extract a feature vector for each term to be categorized we used the British National Corpus [BNC-Consortium, 2000]. The BNC is a very large (over 100 million words) balanced corpus of modern English, both spoken

and written. According to the methodology reported in section 3, we obtained a term-by-document matrix from the BNC, and the we performed a SVD operation on it, to get an LSI space. For each text in the BNC corpus, pos tagging has been performed, multiwords have been identified and sentences have been splitted before collecting the term-by-document matrix. In addition long texts have been divided into into subparts of ten sentences each, and a set 40000 randomly selected sub-documents has been used to perform SVD. From the output of the SVD process we obtained LSI vectors for each lemma contained in WordNet whose document frequency in the BNC corpus is higher than 3. Table 1 describes the lexicon extracted by reporting the number of different lemmata (w), the number of monosemous lemmas (MS), the number of lemmata for which WordNet Domains reports only one non generic (i.e. not Factotum) domain label (MD), and the intersection between MS and MD. We disaggregated the results by POS, and we evaluated this figures for lemmata having a document frequency higher than 10, 50, 100 and 500.

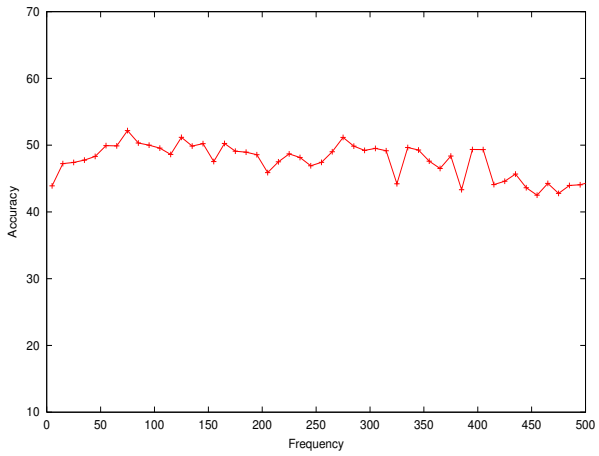

Figure 1: Classification accuracy

5 Results and Discussion

As a preliminary experiment we investigated about the minimum frequency required for a term to be classified, the results are reported in figure 1. Results clearly show that the classification accuracy is pretty constant when the number of occurrences of the term is higher than 10, while it is lower for very rare terms. For the rest of the experiments we used all the lemmata having frequency higher than 10.

We report evaluation results in Table 2. The use of Latent semantic kernels is effective to perform TC, getting an F1 micro measure of about 43%. Even if the results are not directly comparable to those reported in [Lavelli *et al.*, 2002], (e.g. the corpus used for learning was Reuters instead of BNC, ...), the authors reported an F1 micro measure of about 10 - 20% using the same domain set.

Domain	*Prec*	*Rec*	*F1*
music	0.76	0.53	0.63
mathematics	0.70	0.53	0.60
economy	0.55	0.63	0.58
doctrines	0.51	0.65	0.57
chemistry	0.56	0.52	0.54
engineering	0.53	0.53	0.53
alimentation	0.48	0.59	0.53
publishing	0.90	0.37	0.53
pedagogy	0.63	0.44	0.52
commerce	0.58	0.45	0.51
medicine	0.53	0.41	0.46
architecture	0.39	0.55	0.46
law	0.46	0.46	0.46
art	0.47	0.44	0.45
military	0.46	0.42	0.44
zoology	0.42	0.44	0.43
literature	0.64	0.32	0.42
transport	0.49	0.36	0.41
geography	0.33	0.57	0.41
psychology	0.42	0.40	0.41
sport	0.38	0.45	0.41
biology	0.36	0.44	0.40
linguistics	0.37	0.38	0.37
history	0.33	0.40	0.36
social_science	0.26	0.49	0.34
earth	0.42	0.27	0.33
physics	0.42	0.27	0.33
agriculture	0.67	0.20	0.31
politics	0.23	0.38	0.29
anthropology	0.50	0.11	0.18
telecommunication	0.17	0.05	0.08
industry	0.33	0.03	0.06
sociology	0.17	0.03	0.05
administration	0.0	0.0	0.0
Micro	43.3	43.3	43.3

Table 2: TC evaluation for each domain for lemmata having frequency higher than 10

Table 3 reports the misclassification errors made by the TC algorithm, derived from the confusion matrix[5]. It is possible to note that most of the errors are motivated by the existence of "domain families" (i.e. domains whose distinctions and delimitations are not clear) such as Biology - Zoology, Linguistics - Literature, Politics - Administration - Law. Thus most of the errors are "shallow" misclassifications, that could be corrected by simply selecting a more distinguishable set of domain for training, preserving the generality of the technique.

[5]Only domain pairs with a high number of errors are reported in Table 3.

Misclassification errors		*# of errors*
biology	medicine	18
zoology	biology	15
biology	chemistry	14
biology	zoology	13
medicine	biology	13
architecture	transport	12
architecture	biology	9
economy	industry	9
politics	administration	9
biology	physics	8
economy	politics	8
social_science	biology	8
alimentation	biology	7
social_science	physics	7
biology	alimentation	7
geography	earth	7
law	sociology	7
politics	economy	7
biology	earth	6
architecture	industry	6
linguistics	literature	6
social_science	architecture	6
economy	law	6
architecture	economy	6
medicine	psychology	6
architecture	social_science	6
architecture	alimentation	6
commerce	economy	6
politics	military	6
geography	military	5
alimentation	medicine	5
geography	zoology	5
law	medicine	5
psychology	medicine	5
alimentation	architecture	5
social_science	alimentation	5
chemistry	earth	5
biology	transport	5
architecture	art	5
economy	sociology	5
social_science	economy	5
politics	law	5
social_science	psychology	5
...	...	...

Table 3: Misclassification errors among domain pairs

6 Conclusions and Future Work

In this paper we summarize and conclude the research activity in the area of lexical acquisition performed in the framework of the meaning project. We proposed a novel methodology to approach the Term Categorization task, consisting on adopting Latent Semantic Kernels. We evaluated this technique on the task of assigning domain labels to monosemous words. Results show that the proposed technique is effective, achieving an accuracy of about 43% for all the monosemus terms in a corpus. Even though our results are not strictly comparable to those reported in the literature, our technique sensibly improves the state of the art by solving the sparseness problem.

We also reported an error analysis showing that most of the misclassification errors are related the the fuzzy nature of domain distinctions. In particular we identified a set of "families" in the WordNet Domains categories that makes the classification problem arbitrary both for the automatic algorithm and for human annotators.

For the future, we plan to continue our research activity in lexical acquisition even if the MEANING project is concluded. In particular we plan to apply the proposed TC techniques to acquire domain specific terminology from corpora for ontology population, by combining it to fine grained Named Entitiy recognition techniques. In addition we are going to exploit TC to identify domain specific substructures of WordNet, in order to tune the lexical resource for specific domains defined by sets of domain specific words. The use of TC for monosemous terms allows us to associate domain labels to WordNet synset, that can be easily propagated through the WordNet structure.

Acknowledgements

This work was supported by the *Meaning* European Project (IST-200134460). We would like to thank Alberto Lavelli for useful suggestions and discussions.

References

[Avancini *et al.*, 2003] H. Avancini, A. Lavelli, B. Magnini, F. Sebastiani, and R. Zanoli. Expanding domain-specific lexicons by term categorization. In *Proceedings of the 2003 ACM symposium on Applied computing*, pages 793 – 797, 2003.

[Berry *et al.*, 1995] M.W. Berry, S.T. Dumais, and T.A. Letsche. Computational methods for intelligent information access. In *Proceedings of Supercomputing 1995,*, 1995.

[BNC-Consortium, 2000] BNC-Consortium. British national corpus. http://www.hcu.ox.ac.uk/bnc/, 2000.

[Comaroni *et al.*, 1989] J. P. Comaroni, J. Beall, W. E. Matthews, and G. R. New, editors. *Dewey Decimal Classification and Relative Index*. Forest Press, Albany, New York, 20^{th} edition, 1989.

[Deerwester *et al.*, 1990] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 1990.

[Fellbaum, 1998] C. Fellbaum. *WordNet. An Electronic Lexical Database*. The MIT Press, 1998.

[Gliozzo *et al.*, 2004] A. Gliozzo, C. Strapparava, and I. Dagan. Unsupervised and supervised exploitation of semantic domains in lexical disambiguation. *Computer Speech and Language*, 18(3):275–299, July 2004.

[Kearns and Vazirani, 1994] M.J. Kearns and U.V. Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, 1994.

[Lavelli *et al.*, 2002] A. Lavelli, B. Magnini, and F. Sebastiani. Building thematic lexical resources by term categorization. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 2002.

[Magnini and Cavaglià, 2000] B. Magnini and G. Cavaglià. Integrating subject field codes into WordNet. In *Proceedings of LREC-2000*, Athens, Greece, June 2000.

[Magnini *et al.*, 2002] B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8(4):359–373, 2002.