Semantic Domains and Supersense Tagging for Domain-Specific Ontology Learning

Davide Picca, Alfio Massimiliano Gliozzo, Massimiliano Ciaramita

University of Lausanne - CH-1015 Lausanne - Switzerland
davide.picca@unil.ch

Fondazione Bruno Kessler - via Sommarive 18, 38050 Povo (TN) Italy
gliozzo@itc.it

Yahoo! Research Barcelona Ocata 1 08003 Barcelona - Spain
massi@yahoo-inc.com

Abstract

In this paper we propose a novel unsupervised approach to learning domain-specific ontologies from large open-domain text collections. The method is based on the joint exploitation of Semantic Domains and Super Sense Tagging for Information Retrieval tasks. Our approach is able to retrieve domain specific terms and concepts while associating them with a set of high level ontological types, named supersenses, providing flat ontologies characterized by very high accuracy and pertinence to the domain.

1 Introduction

In the Semantic Web paradigm it is required to provide a structured view of the unstructured information expressed in texts. Structured information about a specific domain is in general represented by means of ontologies describing the domain, i.e. an explicit representation of the knowledge shared by a community. The ontology building process is typically performed manually by domain experts, making this approach unrealistic for large corpora. Hence, the problem of automatically acquiring concepts and relations describing a particular domain and populating the derived semantic network of relevant entities and instances, i.e. the Ontology Learning problem [Buitelaar et al., 2005], has become an important subject in Information Retrieval (IR). Natural language processing (NLP) techniques can support the ontology learning process by integrating automatic systems for terminology extraction, word sense disambiguation, and relation extraction.
The main contribution of this paper to the problem of ontology learning is a novel method for automatically acquiring and populating domain specific ontologies from large open-domain text collections. In particular, our system retrieves coarse grained ontologies, composed by simple one-layer associations among domain specific concepts, entities and their ontological type (i.e. the WordNet super senses, such as “artifact”, “act” and “person”), as illustrated in Table 3.
Our method is based on a combination of two basic approaches: (i) Super Sense Tagging (SST) and (ii) Domain Modeling (DM). SST is the problem to identify terms in texts, assigning a "supersense" category (e.g. `person`, `act`) to their senses in context. The hypothesis that we investigate in this paper is that the information provided by supersenses, although fairly coarse-grained and noisy, when paired with domain information can produce quite precise semantic representations. This is a consequence of the fact that the semantic level of representation captured by domains, although coarse-grained as well, is orthogonal to the semantic representation provided by supersenses. Thus, their combination can produce a sort of second-order semantic representations which are able to capture informative semantic aspects of terms.
We adopt SST as a preprocessing step (see Section 2), and we apply it to recognize terms and entities in large collections of texts. Then we perform a distributional analysis of

the occurrences of such terms in the corpus, with the goal of finding domain relations among them (see Section 3). The result of such analysis, that we call Domain Modeling, is a similarity metric among terms and texts, that can be used to query the corpus for domain specific terminology. As a final step, in Section 4 we assigned the more appropriate ontological type to each term, by simply selecting the most frequent supersense in which the term appeared in the domain specific texts, achieving the desirable effect of avoiding the noise due to the tagger.
As illustrated in Section 5, the proposed approach achieves impressive results, as far as the pertinence to the domain and the accuracy of the ontological type recognition phases are concerned, offering an innovative approach to the ontology learning field.

2 Supersense Tagging

WordNet [Fellbaum, 1998] defines 41 lexicographer's categories, also called *supersenses* [Ciaramita and Johnson, 2003], used by lexicographers to provide an initial broad classification for the lexicon entries[1]. Although simplistic in many ways, the supersense ontology has several attractive features for NLP purposes. First, concepts, although fairly general, are easily recognizable. Secondly, the small number of classes makes it possible to implement state of the art methods, such as sequence taggers, to annotate text with supersenses. Finally, similar word senses tend to be merged together. Hence, while the noun *folk* has four fine-grained senses, at the supersense level it only has two as illustrated below:

1. people in general (noun.group)
2. a social division of (usually preliterate) people (noun.group)
3. people descended from a common ancestor (noun.group)
4. the traditional and typically anonymous music that is an expression of the life of people in a community (noun.communication)

Previous work has showed that supersenses can be useful in lexical acquisition to provide a first guess at the meaning of novel words [Ciaramita and Johnson, 2003], and in syntactic parse re-ranking, to define latent semantic features [Koo and Collins, 2005]. Using the Semcor corpus, a fraction of the Brown corpus annotated with WordNet word senses, a supersense tagger has been implemented [Ciaramita and Altun, 2006] which can be used for annotating large collections of English text [2]. The tagger implements a Hidden Markov Model, trained with the perceptron algorithm introduced in [Collins, 2002]. The tagset used by the tagger defines 26 supersense labels for nouns and 15 supersense labels for verbs. The tagger outputs named entity information, but also covers other relevant categories and attempts lexical disambiguation at the supersense level. The following is a sample output of the tagger:

(1) $\text{Guns}_{B-noun.group}$ $\text{and}_{I-noun.group}$ $\text{Roses}_{I-noun.group}$ $\text{plays}_{B-verb.communication}$ at_{O} the_{O} $\text{stadium}_{B-noun.location}$

Compared to other semantic tagsets, supersenses have the advantage of being designed to cover all possible open class words. Thus, in principle, there is a supersense category for each word, known or novel. Additionally, no distinction is made between proper and common nouns, whereas the named entity tag set tends to be biased towards the former.

3 Exploiting Semantic Domains for Ontology Learning

Semantic Domains are common areas of human discussion, such as Economics, Politics, Law [Gliozzo, 2005]. Semantic Domains can be described by DMs [Gliozzo, 2005], by

[1] Throughout the paper we intend WordNet version 2.0.

[2] The tagger is publicly available at: http://sourceforge.net/projects/supersensetag/.

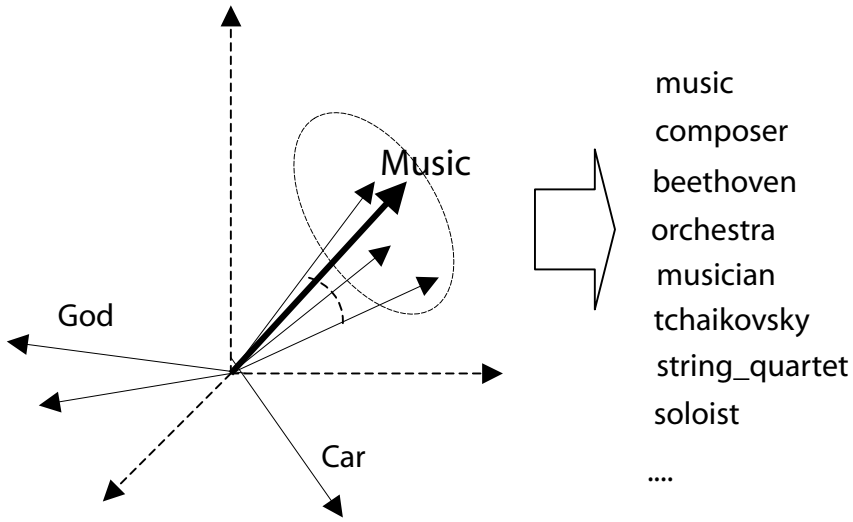

Figure 1: Semantic Domain generated for the query *music*

defining a set of term clusters, each representing a Semantic Domain, i.e. a set of terms having similar topics. A DM is represented by a $k \times k'$ rectangular matrix $\mathbf{D}$, containing the domain relevance for each term with respect to each domain.
DMs can be acquired from texts by exploiting term clustering algorithms. The degree of association among terms and clusters, estimated by the learning algorithm, provides a domain relevance function. For our experiments we adopted a clustering strategy based on Latent Semantic Analysis (LSA) [Deerwester et al., 1990], following the methodology described in [Gliozzo, 2005].
The input of the LSA process is a Term by Document matrix $\mathbf{T}$ of the frequencies in the whole corpus for each term. In this work we indexed all those lemmatized terms recognized by the SST, filtering out verbs. The so obtained matrix is then decomposed by means of a Singular Value Decomposition, identifying the principal components of $\mathbf{T}$.
Once a DM has been defined by the matrix $\mathbf{D}$, the Domain Space is a k' dimensional space, in which both texts and terms are associated to Domain Vectors (DVs), i.e. vectors representing their domain relevance with respect to each domain. The DV $\vec{t_i'}$ for the term $t_i \in \mathcal{V}$ is the i^{th} row of $\mathbf{D}$, where $\mathcal{V} = \{t_1, t_2, \ldots, t_k\}$ is the vocabulary of the corpus. The DVs for texts are obtained by mapping the document vectors $\vec{d_j}$, represented in the vector space model, into the vectors $\vec{d_j'}$ in the Domain Space, defined by

$$\mathcal{D}(\vec{d_j}) = \vec{d_j}(\mathbf{I^{IDF}D}) = \vec{d_j'} \tag{2}$$

where $\mathbf{I^{IDF}}$ is a diagonal matrix such that $i_{i,i}^{IDF} = IDF(w_i)$ and $IDF(w_i)$ is the *Inverse Document Frequency* of w_i. The similarity among both texts and terms in the Domain Space is then estimated by the cosine operation.
When a query Q is formulated, our algorithm retrieve the couple of ranked lists $dom(Q) = (t_1, t_2, \ldots, t_{k_1}), (d_1, d_2, \ldots, d_{k_2})$ of domain specific terms such that $sim(t_i, Q) > \theta_t$ and $sim(d_i, Q) > \theta_d$, where $sim(Q, t)$ is a similarity function capturing domain proximity and θ_t and θ_d are the the *domain specificity* thresholds for terms and texts, respectively. The process is illustrated by Figure 1 . The output of the Terminology Extraction step is then

a ranked list of domain specific candidate terms.

4 Ontological Type Recognition

Our method combines the information provided by the SST and DM in order to reduce the noise of both models and create more complex domain-specific semantic representations. The method works as follows. We use SST to organize the output of the domain model and create a first coarse-grained hierarchy of the domain-specific terminology returned by the domain modeling described in the previous section, identifying groups of concepts and entities belonging to the same ontological type (e.g. `person`, `act`, `group`). However, a certain degree of ambiguity is still present in the list returned by the previous step. In fact, the same term can be annotated by the SST with different supersenses in different contexts. E.g., the term *rock* is both a kind of `communication`, in the "musical_gender" sense, and a kind of `material`, depending on its actual sense. Nevertheless, ambiguity should be solved in a domain specific ontology; e.g., an ontology of the musical domain is expected to contain only the `communication` sense of rock. The disambiguation accuracy of the tagger for each individual token is not good enough for ontology learning, where high degree of precision is necessary. Therefore a further disambiguation step is required, whose aim is to discard noisy sense assignments and to select only domain specific senses of terms.
To address this issue, for each term, we determine the frequency of all its possible supersense assignments in the domain specific collection of documents retrieved in the DM phase, as predicted by SST. Hence, we assign to each term its most frequent supersense, to determine its ontological type. This simple strategy allows us to filter out the noise present in the individual supersense assignments, and to select the most appropriate ontological type for each term in the domain specified by the query.
As an example, the noun "piano" occurs 310 times in music domain texts as a `communication`, and 37 times as a `person`. In such cases, the most frequent strategy filters out the unwanted noisy assignments (`piano/person`). The most frequent strategy provides a good approximation of the most important ontological type of each domain term.
Both supersense tagging and domain analysis can be performed on large scale corpora without requiring any manual intervention. In addition, the flexibility and efficiency of both methods allows us to work with very large corpora, opening an interesting research direction on ontology-based information retrieval.

5 Evaluation

To evaluate the Ontology Learning process described in the previous section we adopted a large open domain text collection and we selected a set of domains by formulating appropriate queries. In this section we first describe the corpora and the tools adopted to implement our algorithms, then we evaluate the quality of the retrieved ontologies in terms of pertinence to the domain and accuracy in the Ontological Type assignments.

5.1 Experimental Settings

In our experiments we used the British National Corpus. We split each text into subportions of 40 sentences, and regarded each portion as a different document, collecting overall about 130,000 documents. Each document was annotated with the supersense tagger. A term by document matrix describing the whole corpus was extracted, where the terms adopted are in the form `term#supersense`, as for example `radio#artifact`. To filter out less reliable low-frequency terms, we considered only those terms occurring in more than 3 documents in the corpus, obtaining a vocabulary of about 450,000 terms. The singular value decomposition (SVD) process was performed by considering the first 100 dimension. This step took about two hours on a laptop with 1GB of memory.

	Person	Cogn	Comm	Act	Event	Artifact	Others
Sport	27.74	0.00	0.73	10.95	2.10	2.10	56.38
Religion	35.00	13.37	8.69	8.36	0.66	1.33	67.60
Music	51.42	1.06	10.00	2.84	1.42	6.76	29.34

Table 1: Percentage of extracted ontological types

	Pertinence	Ontological Type	Number of Terms
Sport	93.15%	89.40%	73
Religion	81.30%	96.00%	299
Music	90.39%	87.90%	281

Table 2: Accuracy of the system.

5.2 Accuracy and Pertinence

We submitted the system three different queries, describing the domains of music, religion and sport, respectively by formulating the queries `Music`, `Religion` and `Sport`. In order to perform this step the empirical thresholds θ_d and θ_t have been empirically set to *0.4* and *0.6*, respectively for documents and terms, observing that these assignments provide good quality domain specific material for any query.

As a result the system provides two ranked lists of domain specific terms and documents. We considered only those ontological types occurring more than 3 times in the domain specific documents, obtaining a total of 300 terms for the domain `Religion`, 73 for the domain `Sport` and 281 for the domain `Music`. From this list, we solved the cases of ambiguous supersense assignments by selecting the most frequent ontological types. As a result we obtained a list of concepts and entities for each class, as illustrated in Table 3. Such an output can be interpreted as a flat (i.e. one layer) ontology describing the domain of the query. Overall, the distribution of the retrieved concepts and entities with respect to their ontological type is reported in Table 1.

Systems for ontology learning are complicated to be evaluated in terms of recall. This problem is even more relevant in an open-domain perspective, where it is impossible to have a clear picture of the domain knowledge actually contained in texts. Therefore, we concentrated on evaluating the accuracy of our system.

To this aim, we submitted the lists of terms retrieved by the system for each query to domain experts, and we asked a lexicographer to judge each term with respect to two perspectives: Pertinence to the domain of the query, and correctness of the Ontological Type assigned. Table 3 summarizes an example of the annotation we did for the domain `Music`. The term "gig" has not been correctly classified by SST (marked as 0 in the column) as `artifact` but it is pertinent to the domain `Music` (marked as 1 in the column). Inversely, the term "vocals" is really pertinent to domain but it is not correctly recognized by the SST.

The overall results are reported in Table 2, showing that the system is highly accurate and able to retrieve domain specific entities and concepts. In particular, the pertinence of the retrieved ontology for the domain `Sport` has the highest value (about 93% of the retrieved terms have been judged pertinent with respect to the domain of the query), while the ontological type is disambiguated best in the domain `Religion` (accuracy 96%). Interestingly, our method can also be used for ontology population because named entities are typically assigned the correct ontological type. For example, in the domain `Sport`, the system extracted *boris_becker*, *monica_seles*, *jim_courier* and assigned the ontological type person to them. As reported in Table 1, most of the extracted concepts and entites belongs to the ontological type `person`. All proper names not existing in Wordnet, have

Artifact	P	O	F	Commun.	P	O	F	Person	P	O	F
recording	1	1	833	music	1	1	2,835	composer	1	1	405
gig	1	0	467	song	1	1	1,620	vocals	1	0	95
disc	1	1	400	story	0	1	313	young	0	1	88
recording_studio	1	1	23	pop_music	1	1	76	Johnny_Marr	1	1	70

Table 3: System output and evaluation for the domain Music. P, O and F indicate the domain Pertinence judgment (boolean), the appropriateness of the Ontological type (boolean) and the Frequency in the domain specific texts.

been correctly disambiguated with a precision of 100%.

6 Conclusion and future work

In this paper we presented a novel approach for ontology learning from open domain text collections, based on the combination of Super Sense Tagging and Domain Modeling techniques. The system recognizes terms pertinent to the domain and assign then the correct ontological type roughly 90% of the time. For the future, we plan to evaluate the system in a more systematic way, by comparing its output to hand-made reference ontologies. To improve the coverage of the system, we are planning to train on a WEB scale text collection. In addition, we plan to provide a fine grained structure to the coarse grained one-layer ontologies presented in this paper, by adopting automatic techniques to identify is_a relations among the retrieved terms, and by distinguishing automatically between concepts and entities. Finally, we plan to explore the use of our methodology to provide additional knowledge to NLP systems for Question Answering, Information Extraction and Textual Entailment.

Acknowledgments

Alfio Gliozzo was supported by the FIRB-Israel co-founded project N.RBIN045PXH.

References

[Buitelaar et al., 2005] Buitelaar, P., Cimiano, P., and Magnini, B. (2005). *Ontology learning from texts: methods, evaluation and applications*. IOS Press.

[Ciaramita and Altun, 2006] Ciaramita, M. and Altun, Y. (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of EMNLP-06*, pages 594–602, Sydney, Australia.

[Ciaramita and Johnson, 2003] Ciaramita, M. and Johnson, M. (2003). Supersense tagging of unknown nouns in wordnet. In *Proceedings of EMNLP-03*, pages 168–175, Sapporo, Japan.

[Collins, 2002] Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP-02*.

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

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

[Gliozzo, 2005] Gliozzo, A. (2005). *Semantic Domains in Computational Linguistics*. PhD thesis, University of Trento.

[Koo and Collins, 2005] Koo, T. and Collins, M. (2005). Hidden-variable models for discriminative reranking. In *Proceedings of EMNLP-05*, Vancouver, Canada.