Harvesting Ontologies from Open Domain Corpora: a Dynamic Approach

R. Basili(*), A. Gliozzo (ℵ), M. Pennacchiotti (‡)
(*) DISP - University of Roma, *Tor Vergata*
Via del Politecnico, 1 - 00133 Roma (Italy)
basili@info.uniroma2.it
(ℵ) Fondazione Bruno Kessler
Povo, Trento (Italy)
gliozzo@itc.it
(‡) Computational Linguistics, Saarland University
Saarbrucken, Germany.
pennacchiotti@coli.uni-sb.de

Abstract

In this work we present a robust approach for dynamically harvesting domain knowledge from open domain corpora and lexical resources. It relies on the notion of Semantic Domains and provides a fully unsupervised method for terminology extraction and ontology learning. It makes use of an algorithm based on Conceptual Density to extract useful relations from WordNet. The method is efficient, accurate and widely applicable, as the reported experiments show, opening the way for effective applications in retrieval tasks and ontology engineering.

Keywords

Lexical Acquisition, Ontology Learning, Word Sense Disambiguation

1 Introduction

Ontology learning from text is a popular field of research in Natural Language Processing (NLP). The increasing amount of textual information at our disposal needs to be properly identified, structured and formalized to make it accessible and usable in applications. Much work has focused on the harvesting phase of ontology learning. Researchers have successfully induced terminologies, word similarity lists [13], generic and domain relations [20, 17], facts [6], entailments [22] and other resources.
However, these resources must be structured in a richer semantic network in order to be used in inference and applications. So far, this issue has been solved by linking the harvested resources into existing ontologies or structured lexical repositories like WordNet [7], as in [16, 21].
Yet, applications often require domain specific knowledge but this means that adapting the existing general purpose resources, such as WordNet, is required. In general, this task is not trivial, as large scale resources are ambiguous (i.e. terms may refer to multiple concepts in an ontology, even if only some of them are actually relevant for the domain) and not balanced (i.e. some portions of WordNet are much more densely populated than others [1]). These problems are typically addressed by performing the following tasks.

Lexical ambiguity resolution : disambiguate terms by linking them to the correct sense(s) for the specific domain.

Ontology pruning : prune the ontology and induce only the sub-portion which is relevant for the given domain. This can be intended as a side effect of ambiguity resolution.

Ontology Population : extend an existing ontology with novel instances, concepts and relations found into domain specific corpora.

Most of these domain-oriented approaches (e.g. [23]) require domain specific corpora and are typically semi-supervised, as they need manual intervention to alleviate the errors due to the typically low precision achieved by automatic techniques. This constraint prevents the use of such techniques into open domain scenarios in applications in which the domain of interest is specified at run-time (such as Information Retrieval (IR) and Question Answering).

In this paper, we propose a solution to the above issue, by focusing on the problem of on-line domain adaptation of large scale lexical ontologies. The requirement for such an application is to implement an adaptation process which is:

- performed at run time;
- tuned by using only the user information need;
- fully automatized, and therefore accurate enough for the application in which it is located.

In contrast to classical approaches, we propose a novel unsupervised technique to induce *on-the-fly* domain specific knowledge from *open domain corpora*, starting from a simple user query formulated in a IR style.

Our algorithm is inspired by the notion of *Semantic Domains* and is based on the combined exploitation of two very well known techniques in NLP: Latent Semantic Analysis (LSA) [5] and Conceptual Density (CD) [1]. The main idea is to first apply LSA to extract a domain terminology from a large open domain corpus, as an answer to the user query. Then, the algorithm leverages CD to project the inferred terms into WordNet to identify domain specific sub-regions in it, that can be regarded as lexicalized core ontologies for the domain of interest. The overall approach allows to achieve the goals of *lexical ambiguity resolution* and *ontology pruning*, and offers an online solution to the problem of domain adaptation of lexical resources discussed in [18, 24]. An example of the output of our system for the query MUSIC is illustrated in Figure 1.

In our setting, the use of LSA guarantees a major advantage. Unlike classical methods to estimate term similarity (e.g. [25, 12]) which are based on contextual similarity [4], LSA relies on a domain restriction hypothesis [10] stating that two terms are similar, and therefore are very likely to be semantically related, when they belong to the same domain, i.e. when they co-occur in the same texts. LSA detects as similar terms not those having the same ontological type (e.g. the most similar terms to *doctor* will be concepts belonging to the type *PERSON*) but those referring to the same domain, as needed in ontology learning (for example, in the medical domain we need both *doctors*, and *hospital*).

In the rest of the paper we will show evidences supporting the following contributions of this work: (i) the induction process is triggered by a simple IR-like query, providing to the user/application the required domain ontology *on the fly*; (ii) unlike previous approaches, our method does not need domain corpora, (iii) the method guarantees high precision both in the lexical ambiguity resolution and in the ontology induction phases.

We will also show that the main contribution of our method is a very accurate Word Sense Disambiguation (WSD) algorithm, largely outperforming a most frequent baseline and achieving performance close to human agreement. The paper is organized as follows. In Section 2 we introduce the concept of Semantic Domain as a theoretical framework motivating our work and we describe the terminology extraction step, required to provide an input to the CD algorithm producing the final domain ontology (Section 3). Section 4 concerns evaluation issues, while Section 5 concludes the paper.

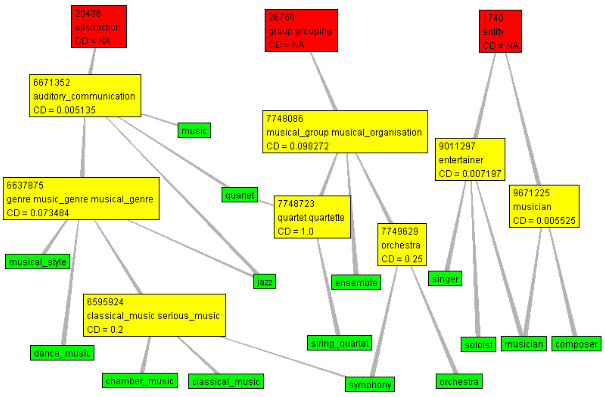

Fig. 1: *Core ontology extracted from WordNet for the “**music**” domain*

2 Terminology Extraction in the Domain Space

The theoretical foundation underlying this work is the concept of *Semantic Domain*, introduced for WSD purposes [14] and further exploited in different tasks, such as Text Categorization and Relation Extraction [8]. Semantic Domains are common areas of human discussion, such as Economics, Politics and Law. Three properties of Semantic Domains are relevant for our task. First, they are characterized by high lexical coherence [14]. This allows us to automatically induce specific terminologies from open domain corpora. Secondly, the ambiguity of terms in specific domains decreases drastically, motivating our lexical ambiguity resolution process. For example, the (potentially ambiguous) word *virus* is fully disambiguated by the domain context in which it is located (it is a *software agent* in the COMPUTER SCIENCE domain and a *infectious agent* in the MEDICINE domain). Third, as shown in [8], semantic relations tend to be established mainly among domain specific terms.

Semantic Domains are described by Domain Models (DM) [9], by defining a set of term clusters, each representing a Semantic Domain, i.e. a set of terms having similar topics (see Figure 2). DMs can be acquired from texts by exploiting term clustering algorithms. For our experiments we adopted a clustering strategy based on LSA, following the methodology described in [9].

To this aim, we first identify candidate terms in the open domain document collection by imposing simple regular expressions on the output of a Part of Speech tagger (e.g. ((Adj|Noun)+|((Adj|Noun)*(NounPrep)?)(Adj|Noun)*)Noun)), as described in [11]. The obtained term by document matrix is then decomposed by means of Singular Value Decomposition (SVD) [5] in a lower dimensional domain matrix $\mathbf{D}$. The i^{th} row of $\mathbf{D}$ represents the Domain Vector (DV) for the term $t_i \in \mathcal{V}$, where $\mathcal{V} = \{t_1, t_2, \ldots, t_k\}$ is the vocabulary of the corpus (i.e.,the terminology). DVs represent the domain relevance of both terms and documents with respect to any domain. $\mathbf{D}$ is then used to estimate the similarity in a Domain Space (i.e. a k' dimensional space in which both documents and terms are associated to DVs) by using the cosine operator on the DVs.

When a query Q is formulated (e.g. MUSIC), our algorithm retrieves the ranked list $dom(Q) = (t_1, t_2, \ldots, t_{k_1})$ of domain specific terms such that $sim(t_i, Q) > \theta$ where $sim(Q, t)$ is the cosine between the DVs corresponding to Q and t, capturing domain proximity, and θ_t is the *domain specificity* threshold.

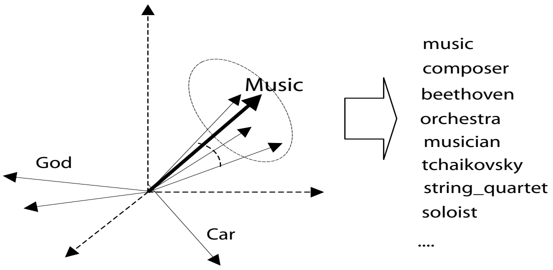

Fig. 2: *Semantic Domain generated by the query* MUSIC

The process is illustrated in Figure 2. The output of the Terminology Extraction step is then a ranked list of domain specific candidate terms and an associate ranked list of domain specific documents.

3 Inducing a core ontology via Conceptual Density

Once a semantic domain has been identified as an unstructured set of domain specific terms, our algorithm induces a core ontology from WordNet, by selecting the maximally dense sub-regions including them. This step involves a WSD process, as only the domain specific synsets associated to the terms extracted in the previous step have to be selected. To induce the core ontology from the terminology, we developed an algorithm, based on CD, that adapts the Dynamic Domain Sense Tagging algorithm proposed in [2]. The goal of our algorithm is twofold:

1. *Lexical ambiguity resolution*. Selecting the domain specific senses of ambiguous domain specific words.

2. *Ontology induction/pruning*. Selecting the best generalizations of the domain specific concepts associated to the word senses.

The algorithm achieves these goals applying a variant of the notion of CD proposed in [3] In the literature, the classical notion of CD has been applied in "local" context of words to be disambiguated, represented as word sets. The main problem of this approach is that small contexts, typically composed by few words appearing in the same sentence, do not allow generalization over the WordNet structure, being them typically spread in the graph, and then not well connected. For example the words *surgeon* and *hospital* lie in different WordNet hierarchies, preventing us from finding the common generalization necessary for disambiguation via CD.

To solve the problem, we apply the CD definition given in [3], integrating it with Domain Information, as in [2]. The context is here intended as the domain terminology $dom(Q)$ inferred from the previous step. The terminology provides the evidence needed to start the generalization process (e.g. in the medical domain we expect to find much more words related to *surgeon*, such as *oncologist* and *dentist*, both related by the common hyperonym *doctor*).

The hypothesis is that when all the paradigmatic relations among terms in $dom(Q)$ are imposed, the CD algorithm is able to select the proper sub-region of WordNet containing the suitable domain specific concepts, discarding most of irrelevant senses associated to the extracted terminology. The outcome of the process is thus the subset of senses or their generalizations able to explain $dom(Q)$ according to WordNet. The result is a "*view*" of the original WordNet, as the core domain ontology for Q (Figure 1).

Specifically, terms $t \in dom(Q)$ can be generalized through their senses σ_t in the WordNet hierarchy. The likelihood of a sense σ_t is proportional to the number of other terms $t' \in dom(Q)$ that have common generalizations with t along the paths activated by their hyperonyms α in the hierarchy. A measure of the suitability of the synsets α for the terms in $dom(Q)$ is thus the *information density* of the subtrees rooted at α. The higher is the number of nodes under α that generalizes some nouns $t \in dom(Q)$, the better is the interpretation α for $dom(Q)$. The CD of a synset α given a query Q, $cd^Q(\alpha)$, models the former notion and provides a measure for the latter.

Ontology Induction. The target core ontology is the set of synsets $G(Q)$ that represents the *best paradigmatic interpretation* of the domain lexicon $dom(Q)$. This can be efficiently computed by the *greedy* search algorithm described in [3] that outputs the minimal set $G(Q)$ of synsets that are *the maximally dense generalizations of at least two terms* in $dom(Q)$. Terms $t \in dom(Q)$ that do not have a generalization are not represented in $G(Q)$[1].

As any $\alpha \in G(Q)$ is a WordNet sysnset, by completing $G(Q)$ with the topmost nodes we obtain a subset of WordNet that can be intended as a full domain-specific ontology for the triggering domain Q. An excerpt of the core domain ontology, for $Q = \{music\}$ is shown in Figure 1 where terms are leaves (*green nodes*), *yellow nodes* are their common hyperonyms $\alpha \in G(Q)$ and *red nodes* are the topmost nodes.

The core ontology, triggered by the short specification of a domain of interest given in Q, is thus the comprehensive explanation of all the paradigmatic relations between terms of the same domain.

Lexical ambiguity resolution. The semantic disambiguation of a target term $t \in dom(Q)$ depends on the subset of generalizations $\alpha \in G(Q)$ concerning some of its senses σ_t. Let $G_t(Q)$ be such a subset, i.e.

$$G_t(Q) = \{\alpha \in G(Q) \mid \exists \sigma_t \text{ such that } \sigma_t \prec \alpha\} \quad (1)$$

where $\prec$ denotes the transitive closure of the hyponymy relation in WordNet. The set $\sigma(t, Q)$ of inferred domain specific sense σ_t for t is given by:

$$\sigma(t, Q) = \{\sigma_t \quad | \quad \sigma_t \prec \overline{\alpha}\} \quad (2)$$

where $\overline{\alpha} = argmax_{\alpha \in G_t(Q)} cd^Q(\alpha)$. Also, multiple senses may be assigned to a term. The CD score associated to each inferred domain sense $\sigma_i \in \sigma(t, Q)$ (i.e. $cd^Q(\overline{\alpha}_i)$) is then mapped to the probability $P(\sigma_i|t, Q)$, which accounts for how reliable the sense is for the term t in the given domain, by normalizing them so that their sum over all senses of t is equal to 1.

4 Evaluation

Our evaluation aims at assessing the ability of our model in: (1) determining a suitable terminological lexicons; (2) extracting a proper ontological description of the target domain. We then focus on measuring the precision of the terminology extraction step in proposing correct candidates (Subsection 4.1), and on the accuracy and coverage of the induced core ontology (Subsection 4.2).

4.1 Terminology Extraction

4.1.1 Experimental Settings

We evaluated terminology extraction in 5 different domains: MUSIC, CHEMISTRY, COMPUTER_SCIENCE, SPORT and CINEMA. We described them by simple queries made by their single names (e.g. SPORT is described by the query "*Sport*"). As open domain corpus, we adopted the British National Corpus (BNC). In a preprocessing step, we split texts into 40 sentence segments, regarded as different documents, amounting to about 130,000 documents.

[1] A Web version of the greedy CD-based algorithm ia available at `http://ai-nlp.info.uniroma2.it/Estimator/cd.htm`.

Each document is PoS-tagged and terms are identified by regular expressions as in [11]. Terms occurring in less than 4 documents are filtered out so that a source vocabulary of about 450,000 different terms is obtained. We run the SVD process on the resulting 450,000 x 130,000 term by document matrix, and we induce a DM from it, by considering a cut to the first 100 dimensions[2].

For each domain, we use the similarity function sim (Section 2) to rank the candidate terms thus obtaining a ranked list of the overall dictionary. To carry out the evaluation we extract a sample of candidate terms in different positions in the list. Specifically, we divide the list in 11 rank levels, and extract 20 random terms from each of the level. The samples are then submitted (neglecting the ordering) to two domain experts. Each term is judged as *Relevant* or *Not Relevant* for the query domain or *Errors* for ill formed expressions (e.g. olive_neighbour), unmeaningful (e.g. aunty_yakky_da) or non-terms (e.g. good_music). For each rank level, the percentage of each label over the 20 candidates is computed. Results for the domain MUSIC are reported in Figure 3[3].

4.1.2 Results

As far as recall is concerned, systems for terminology extraction are hard to evaluate [19]. This problem is even more relevant in an open domain scenario, where it is not possible to have a comprehensive picture of the domain knowledge actually contained in texts. Thus we focused only on evaluating precision.

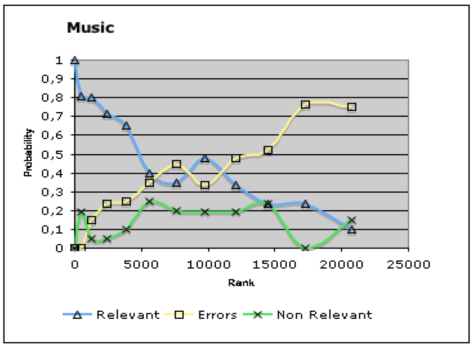

Fig. 3: *Evaluation of the Terminology Extraction algorithm for the* MUSIC *domain*

Results in Figure 3 show that Domain similarity is highly correlated to the precision of the terminology extraction step, providing an effective selection criterion. Setting the domain similarity threshold to 0.8, the algorithm retrieves about 2500 terms, among which 80% are relevant for the domain. When the domain is less represented in the corpus the number of terms retrieved with the same threshold is sensibly lower (e.g. in the domain chemistry the algorithm retrieves about 20 terms), but the accuracy is basically preserved. Therefore domain similarity provides a meaningful selection criterion to retrieve domain specific terminology, ensuring very accurate results without requiring further domain specific parameter settings. We also compared our term extractor to a baseline heuristic, consisting on ranking the same terms with respect to their *frequency* in the top 1,000 domain specific documents for each query, obtained according to their similarity with respect to the initial query (as described in [5]). The precision of the two systems is measured against the labeling of the domain experts of the best ranked 100 terms proposed by each system. Results for all the domains are reported in Table 1. Our algorithm largely outperforms the baseline on all domains.

Domain	TE	Baseline
Chemistry	**0.85**	0.58
Cinema	**0.93**	0.34
Computer	**0.92**	0.46
Music	**0.93**	0.46
Sport	**0.95**	0.48

Table 1: *Precision of our term extractor (TE) and the baseline system, on the top ranked 100 terms for each domain.*

The lower performance obtained on the CHEMISTRY domain are due to the inclusion in the LSA space of some documents/terms relevant for the more general academic domain, which in the BNC slightly overlaps with chemistry. While these are only preliminary results, they show that a LSA based algorithm for ranking terms offers a high degree of precision and can be effectively adopted to perform on-line terminology extraction.

4.2 Inducing Domain Specific Core Ontologies

The goal of the ontology pruning step is to identify coherent sub-portions of WordNet as useful models for a domain: the hypothesis is that these contain most of the selected terms and their generalizations. The CD algorithm presented in Section 3 achieves both goals. In this section we evaluate the ontology pruning step according to two factors: the ability of identifying only correct senses for the terms (Subsection 4.2.2); the "*capacity*" of the core ontologies, i.e. their ability to be populated by novel concepts and/or instances (Subsection 4.2.3).

4.2.1 Experimental Settings

The induction of the core ontology in each area of interest is based on Wordnet (version 2.0). We focused on the noun hierarchy, which is organized on 41 taxonomies describing the hyponymy relation. Due to its huge dimension, pruning WordNet is not an easy task. Out of the 115,524 synsets in WordNet, a core ontology is expected to contain only hundreds of concepts, making the retrieval problem very hard. Given the quality of the terminology extraction process we used as seed the list of domain specific terms for each domain. For each domain we selected all the lemmata in WordNet comprises within the top ranked 1,000 terms for each domain (set r in Section 3) to initialize the CD algorithm. The result is the best (i.e. most conceptually dense) Wordnet substructure. An example is in Figure 1 and 4. Each term that appears in the ontology is also disambiguated, as the CD provides very low scores (close to 0) for all unrelevant senses, which are then discarded in the ontology generation phase.

4.2.2 Identifying domain specific senses

In a first analysis we focused on unambiguous terms, as their corresponding synsets are necessarily domain specific

[2] SVD is applied through LIBSVDC (http://tedlab.mit.edu/∼dr/SVDLIBC/)

[3] Results on other domains do not significantly differ from those reported for Music and will be not reported because of space limitation.

senses. The percentage of monosemous words varies sensibly among the different domains, ranging from 48% in MUSIC to 84% in CHEMISTRY. Figure 3 suggests that less than 20 % of entries within the first 1,000 candidates are not relevant for the ontology. An analysis of the first 200 monosemous terms in the candidate list has been carried out for all domains revealed that about 95% of terms are correct. In such cases the accuracy of the method is higher, as monosemous terms included in Wordnet, are clearly less affected by errors.

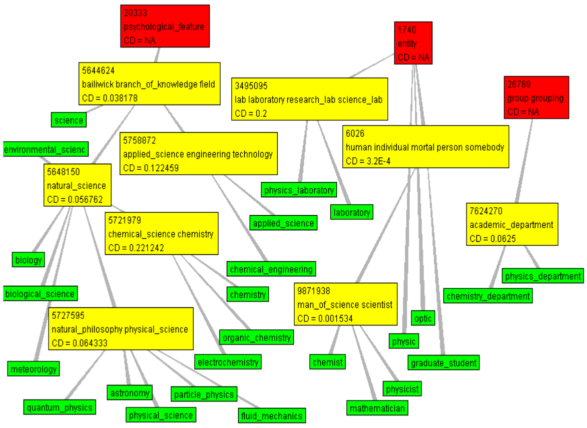

Fig. 4: *Core ontology extracted from WordNet for the* CHEMISTRY *domain*

The real issue is here to validate the senses proposed for ambiguous domain specific terms. This can be regarded as an unsupervised disambiguation task, as we did not use any training data. In contrast to the common WSD settings (where WSD is evaluated as the selection of the correct sense for words in a textual context), we need to measure the ability of selecting *domain specific senses*. In the literature this problem has been also referred as *predominant sense identification* for specific domains, e.g. [15]. Unlike these approaches, our algorithm does not require domain specific collections nor the use of any complex preprocessing tool (e.g. a dependency parser).

To evaluate the disambiguation accuracy, we selected from the top 200 terms in the ranked list of each domain all the ambiguous terms contained in WordNet. We then asked two lexicographers to mark their senses with respect to the query: domain vs. non-domain specific senses are thus labeled. For example, the lemma *percussion* has four senses (i.e. *"the act of playing a percussion instrument"*, *detonation*, *rhythm_section* and *pleximetry*), but only the first and the third have been judged relevant for the domain MUSIC. Table 2 shows some statistics about the annotated resource produced as a gold standard. For each domain, the number of ambiguous cases analyzed and the relative polisemy (according to Wordnet 2.0) is reported in the first two columns. The last two columns report two different inter-annotator agreement measures. *AgrF* represents the "*full*" agreement, estimated by counting all senses in which the annotators agreed (either positives or negatives) and by dividing it by the number of all possible senses. This figure provides an upper bound for the *accuracy* of the system. Since we are mostly interested in defining an upper bound for the F1, we computed a second agreement score. As precision and recall are measured on the positive senses only, the last column (*AgrP*) reports the agreement on positive examples, computed over those cases in which *at least one annotator* provided a positive labeling.

Domain	**Amb**	**Pol**	**AgrF**	**AgrP**
Music	35	3.9	0.91	0.83
Sport	21	5.6	0.92	0.76
Computer	16	4.8	0.97	0.89
Chemistry	9	3.7	0.74	0.53
Cinema	4	5.3	0.95	0.85
Total	95	4.0	0.91	0.78

Table 2: *Domain Specific Gold Standards for Sense disambiguation*

The output of the CD algorithm is an estimation of the probability, for each sense, to be relevant for the domain expressed by the query. We can obtain a flexible binary classifier imposing a threshold $\tau > 0$ on the output sense probabilities: a sense is accepted *iff* its probability is above τ. Figure 5 shows the micro F1, averaged over all domains, obtained by the classifier parameterized with different values of τ, (i.e. from 0, all accepted, to 1, none accepted).

The best F1 value (i.e. 0.75) is obtained by selecting all those senses whose probability is above 0.1. The system is also very precise, at cost of some points of recall: precision is over 0.8 at recall 0.56, and over 0.9 at recall 0.2. This trade-off is interesting as in ontology learning more precise results are often preferable.

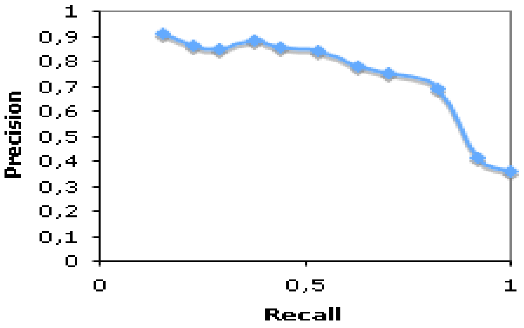

Fig. 5: *Precision and recall for different probability thresholds obtained by the WSD algorithm.*

Table 3 summarizes the individual F1 scores over positive examples, in all domains, obtained with the optimal settings of the classification threshold, i.e. $\tau = 0.1$ [4]. Two different baselines are reported: random and most frequent sense selection. The model outperforms both baselines. Notice how the performance is close to the upper bound provided by the agreement *AgrP* on positive examples of Table 2. As the CD algorithm is fully unsupervised, the improvement on the first sense heuristic is a very good result.

Dom	**Prec**	**Rec**	F1	rnd	MF
Mus	0.85	0.88	**0.87**	0.27	0.38
Spo	0.54	0.71	0.61	0.22	**0.67**
Com	0.58	0.82	**0.68**	0.23	0.18
Chem	0.64	0.875	**0.74**	0.32	0.29
Cine	0.56	0.71	0.63	0.22	**0.72**
Micro	0.69	0.82	**0.75**	0.25	0.40

Table 3: *WSD performances*

[4] Although this setting is derived from the test set itself, it is worthwhile to remark that the same optimal value is preserved over all domains.

4.2.3 Capacity

A final evaluation has been carried out to measure the capability of the core ontologies to host novel concepts and/or instances retrieved in the terminology extraction phase (i.e. their *capacity*). We gave to domain experts the lists of the top ranked 100 terms not included in WordNet for the MUSIC and CHEMISTRY domains. Then, they were asked to judge whether it was possible to attach the terms not in WordNet either to a *High* Level concept in the ontology (i.e. the topmost nodes, such as *entity* or *person*) or to a *domain* specific concept (i.e. the leaves in the ontology). Terms that could not be attached to any node of the core ontology have been marked as *Null*. Results are reported in Table 4. As the class of *Null* terms is also including errors from the terminology acquisition step, we can conclude that most of the terms are covered by the acquired domain ontology and can then be further exploited to populate domain specific nodes.

	NULL	HIGH	DOMAIN
MUSIC	22%	31%	*47%*
CHEMISTRY	46%	7%	*47%*

Table 4: *Capacity evaluation. Percentage of terms not in Wordnet covered by the automatically extracted core ontologies*

5 Conclusions and Future Work

In this paper we proposed a robust and widely applicable approach for dynamically harvesting domain knowledge from general corpora and lexical resources . The method exploits the notion of Domain Space and an n-ary semantic similarity measure over Wordnet for terminology extraction and ontology acquisition. Both processes are very accurate, fully unsupervised and efficient. The disambiguation power of the entire chain is very good, largely outperforming traditional effective baselines. The good impact over complex tasks such as term disambiguation and projection of suitable hyponymy/hyperonymy relations in Wordnet opens a number of potential applications. From a methodological point of view, we plan to extend the acquisition process targeting novel relations among concepts implicitly embodied in the original corpus. Also, we plan to develop automatic methods to further populate the core ontology with novel terms retrieved in the terminology extraction phase. The *on-the-fly* derivation of ontological descriptions for the specific domain of interest can be very attractive in Web applications (e.g. querying or navigation scenarios) and every process dealing with complex (e.g. distributed on-line) meaning negotiation problems. A tool for the automatic compilation of the induced ontology into standard knowledge representation formalisms for the semantic WEB, like OWL, is currently under development, as a general Web service to be easily integrated into an Ontology Engineering framework.

Acknowledgments

All authors are grateful to Marco Cammisa for his technical contribution to the experiments. Alfio Gliozzo was supported by the FIRB-israel co-founded project N.RBIN045PXH.

References

[1] E. Agirre and G. Rigau. Word sense disambiguation using conceptual density. In *Proceedings of COLING-96*, Copenhagen, Denmark, 1996.

[2] R. Basili, M. Cammisa, and A. Gliozzo. Integrating domain and paradigmatic similarity for unsupervised sense tagging. In *In Proceedings of ECAI06*, 2006.

[3] R. Basili, M. Cammisa, and F. Zanzotto. A semantic similarity measure for unsupervised semantic disambiguation. In *Proceedings of LREC-04*, Lisbon, Portugal, 2004.

[4] I. Dagan. *Contextual Word Similarity*, chapter 19, pages 459–476. Mercel Dekker Inc, Handbook of Natural Language Processing, 2000.

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

[6] O. Etzioni, M. Cafarella, D. Downey, A.-M. A.M. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–143, 2005.

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

[8] A. Gliozzo. The god model. In *Proceedings of EACL-2006*, Trento, 2006.

[9] A. Gliozzo, C. Giuliano, and C. Strapparava. Domain kernels for word sense disambiguation. In *Proceedings of ACL-2005*, 2005.

[10] A. Gliozzo, M. Pennacchiotti, and P. Pantel. The domain restriction hypothesis: Relating term similarity and semantic consistency. In *In proceedings of NAACL-HLT-06*, 2006.

[11] J. S. Justeson and S. M. Katz. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 1995.

[12] D. Lin. Automatic retrieval and clustering of similar words. In *COLING-ACL*, pages 768–774, 1998.

[13] D. Lin and P. Pantel. DIRT-discovery of inference rules from text. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD-01)*, San Francisco, CA, 2001.

[14] 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.

[15] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. Finding predominant senses in untagged text. In *Proceedings of ACL-04*, pages 280–287, Barcelona, Spain, 2004.

[16] P. Pantel. Inducing ontological co-occurrence vectors. In *Proceedings ACL-2005*, Ann Arbor, Michigan, June 2005.

[17] P. Pantel and M. Pennacchiotti. Espresso: A bootstrapping algorithm for automatically harvesting semantic relations. In *Proceedings of COLING/ACL-06*, 2006.

[18] B. S. Paul Buitelaar. Ranking and selecting synsets by domain relevance. In *Proceedings on NAACL-2001 Workshop on WordNet and Other Lexical Resources Applications, Extensions and Customizations*, Pittsburgh, USA,, 2001.

[19] M. Pazienza, M. Pennacchiotti, and F. Zanzotto. Terminology extraction: an analysis of linguistic and statistical approaches. In S.Sirmakessis, editor, *Knowledge Mining*, volume 185. Springer Verlag, 2005.

[20] D. Ravichandran and E. Hovy. Learning surface text patterns for a question answering system. In *Proceedings of ACL-02*, 2002.

[21] R. Snow, D. Jurafsky, and A. Ng. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the ACL/COLING-06*, pages 801–808, Sydney, Australia, 2006.

[22] I. Szpektor, H. Tanev, I. Dagan, and B. Coppola. Scaling web-based acquisition of entailment relations. In *Proceedings of EMNLP-2004*, Barcellona, Spain, 2004.

[23] P. Velardi, R. Navigli, A. Cucchiarelli, and F. Neri. *Ontology Learning from Text: Methods, Evaluation and Applications*, chapter Evaluation of OntoLearn, a Methodology for Automatic Learning of Domain Ontologies. IOS Press, 2005.

[24] P. Vossen. Extending, trimming and fusing wordnet for technical documents. In *Proceedings on NAACL-2001 Workshop on WordNet and Other Lexical Resources Applications, Extensions and Customization*, Pittsburgh, USA, 2001.

[25] D. Widdows. *Geometry and Meaning*. CSLI Publications, 2004.