Acquiring Legal Ontologies from Domain-specific Texts

Felice Dell'Orletta[1], *Alessandro Lenci*[2], *Simonetta Montemagni*[1],
Simone Marchi[1], *Vito Pirrelli*[1], *Giulia Venturi*[1]

[1]Istituto di Linguistica Computazionale, CNR, Pisa, Italy
[2]Department of Linguistics, University of Pisa, Italy

Abstract

The paper reports on methodology and preliminary results of a case study in automatically extracting ontological knowledge from Italian legislative texts in the environmental domain. We use a fully-implemented ontology learning system (T2K) that includes a battery of tools for Natural Language Processing (NLP), statistical text analysis and machine language learning. Tools are dynamically integrated to provide an incremental representation of the content of vast repositories of unstructured documents. Evaluated results, however preliminary, are very encouraging, showing the great potential of NLP-powered incremental systems like T2K for accurate large-scale semi–automatic extraction of legal ontologies.

Index Terms: ontology learning, document management, knowledge extraction from texts, Natural Language Processing

1. Introduction

The last few years have witnessed a growing body of research and practice aimed at developing legal ontologies for application in the law domain. A number of legal ontologies have been proposed in a variety of research projects, mostly focusing on upper level concepts hand-crafted by domain experts (see [12], for a recent survey). It goes without saying that realistically large knowledge–based applications in the legal domain will need more and more comprehensive ontologies, incrementally integrating continuously updated knowledge. In this perspective, techniques for automated ontology–learning from texts are expected to play an increasingly more prominent role in the near future.

To our knowledge, however, relatively few attempts have been made so far to automatically induce legal domain ontologies from texts. This is the case, for instance, of [10], [11] and [14].The work illustrated in this paper represents another attempt in this direction. It reports the results of a case study carried out in the legal domain to automatically induce ontological knowledge from texts with an ontology learning system, hereafter referred to as T2K (TexttoKnowledge), jointly designed and developed by the Institute of Computational Linguistics (CNR) and the Department of Linguistics of the University of Pisa. The system offers a battery of tools for Natural Language Processing (NLP), statistical text analysis and machine language learning, which are dynamically integrated to provide an accurate representation of the content of vast repositories of unstructured documents in technical domains (DellOrletta et al., 2006). Text interpretation ranges from acquisition of lexical and terminological resources, to advanced syntax and ontological/conceptual mapping. Interpretation results are annotated as XML metadata, thus offering the further bonus of a growing interoperability with automated content management systems for personalized knowledge profiling. Prototype versions of T2K are currently running on public administration portals and have been used for indexing Elearning and Ecommerce materials. In what follows, we report some ontology learning experiments carried out with T2K on Italian legislative texts.

2. From Text to Knowledge

Technologies in the area of knowledge management and information access are confronted with a typical acquisition paradox. As knowledge is mostly conveyed through text, content access requires understanding the linguistic structures representing content in text at a level of considerable detail. In turn, processing linguistic structures at the depth needed for content understanding presupposes that a considerable amount of domain knowledge is already in place. Structural ambiguities, long-range dependency chains, complex domain-specific terms and the ubiquitous surface variability of phraseological expressions require the operation of a battery of disambiguating constraints, i.e. a set of interface rules mapping the underlying conceptual organization of a domain onto surface language. With no such constraints in place, text becomes a slippery ground of unstructured, strongly perspectivized and combinatorially ambiguous information bits.

There is no simple way around this paradox. Pattern matching techniques allow for fragments of knowledge to be tracked down only in limited text windows, while foundational ontologies turn out to be too general to make successful contact with language variability at large. The only effective solution, we believe, is to face the paradox in its full complexity. An incremental interleaving of robust parsing technology and machine learning techniques can go a long way towards meeting this objective. Language technology offers the jumping–off point for segmenting texts into grammatically meaningful complex units and organizing them into non recursive phrasal "chunks" that require no domain–specific knowledge. In turn, chunked texts can sensibly be accessed and compared for statisticallysignificant patterns of domain-specific terms to be tracked down. Surely, this level of paradigmatic categorization is still very rudimentary: at this stage we do not yet know how chunked units are mutually related in context (i.e. what grammatical relations link the min texts) or how similar they are semantically. To go beyond this stage, we suggest getting back to the syntagmatic organization of texts. Current parsing technologies allow for local dependency relations among chunks to be identified reliably. If a sufficiently large amount of parsed text is provided, local dependencies can be used to acquire a first level of domain-specific conceptual organization. We can then use this preliminary conceptual map for harder and longer dependency chains to be parsed and for larger and deeper conceptual networks to be acquired. To sum up, facing the bootstrapping paradox requires an incremental process of annotation-acquisition-annotation, whereby domain-specific knowledge is acquired from linguistically-annotated texts and then projected back onto texts for extra linguistic information to be annotated and further knowledge layers to be extracted.

To implement this scenario, a few NLP ingredients are required. Preliminary term extraction presupposes postagged texts, where each word form is assigned the contextually appropriate part-of-speech and a set of morpho-syntactic features plus an indication of lemma. Whenever more information about the local syntactic context is to be exploited, it is advisable that basic syntactic structures are identified. As we shall see in more detail below, we use chunking technology to attain this level of basic syntactic structuring. NLP requirements become more demanding when identified terms need be organised into larger conceptual structures and connected through long-distance relational information. For this purpose syntactic information must include identification of dependencies among lexical heads. The approach to ontology learning adopted by T2K exploits all these levels of linguistic annotation of texts in an incremental fashion. Term extraction operates on texts annotated with basic syntactic structures (so–called "chunks"). Identification of conceptual structures, on the other hand, is carried out against a dependency-annotated text.

3. T2K architecture

T2K is a hybrid ontology learning system combining linguistic technologies and statistical techniques. T2K does its job into two basic steps:

1. extraction of domain terminology, both single and multi–word terms, from a document base;
2. organization and structuring of the set of acquired terms into proto–conceptual structures, namely
   - fragments of taxonomical chains, and
   - clusters of semantically related terms.

Figure 1 illustrates the functional architecture of T2K:

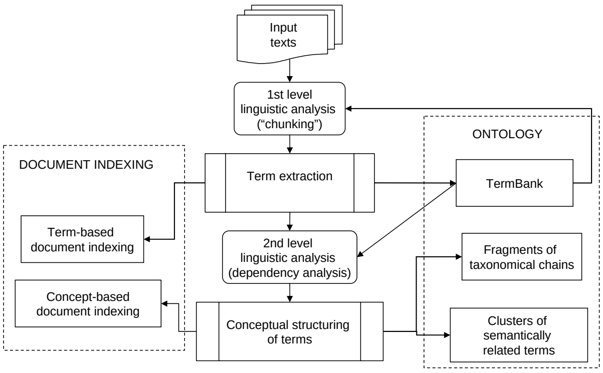
Input texts
1st level linguistic analysis ("chunking")
DOCUMENT INDEXING
ONTOLOGY
Term extraction
TermBank
Term-based document indexing
2nd level linguistic analysis (dependency analysis)
Fragments of taxonomical chains
Concept-based document indexing
Conceptual structuring of terms
Clusters of semantically related terms

Figure 1: *T2K architecture.*

The two basic steps take the central pillar of the portrayed architecture, showing the interleaving of Natural Language Processing (NLP) and statistical tools. Acquired results are structured in the ontology box on the right–hand–side of the diagram, whose stratified organization is reminiscent of the hierarchical cascade of knowledge layers in the "Ontology Learning Layer Cake" by [6], going from terminological information to proto–conceptual structures corresponding to taxonomical and non-hierarchical relationships among terms. Acquired knowledge is also used for document indexing, on the basis of extracted terms and acquired conceptual structures. In what follows we focus on the ontology learning process.

3.1. Term extraction

Term extraction is the first and most–established step in ontology learning from texts. For our present purposes, a term can be a common noun as well as a complex nominal structure with modifiers (typically, adjectival and prepositional modifiers).

T2K looks for terms in shallow parsed texts, i.e. texts segmented into an unstructured (non-recursive) sequence of syntactically organized text units called "chunks" (e.g. nominal, verbal, prepositional chunks). Candidate terms may be one word terms ("single terms") or multi–word terms ("complex terms"). The acquisition strategy differs in the two cases.

Single terms are identified on the basis of frequency counts in the shallow parsed texts, after discounting stop–words. The acquisition of multi–word terms, on the other hand, follows a two–stage strategy. First, the chunked text is searched for on the basis of a set of chunk patterns. Chunk patterns encode syntactic templates of candidate complex terms: for instance, adjectival modification (e.g. *organizzazione internazionale* 'international organisation'), prepositional modification (e.g. *commercializzazione di autovetture* 'marketing of cars'), including more complex cases where different modification types are compounded (e.g. *commercio di prodotti fitosanitari* 'trade of fitosanitary products'). Secondly, the list of acquired potential complex terms is ranked according to their log–likelihood ratio [8], an association measure that quantifies how likely the constituents of a complex term are to occur together in a corpus if they were (in)dipendently distributed, where the (in)dependence hypothesis is estimated with the binomial distribution of their joint and disjoint frequencies.

Recognition of longer terms is carried out by iteratively applying the extraction process to the results of the previous acquisition step. This means that acquired complex terms are projected back onto the original text and the acquisition procedure is iterated on the newly annotated text. The method proves helpful in reducing the number of false positives consisting of more than two chunks [4]. Interestingly, the chunk patterns used for recognition of multi–word terms need not necessarily be the same across different iteration stages. In fact, it is advisable to introduce potentially noisy patterns (such as, for example, coordination patterns) only at later stages.

The iterative process of term acquisition yields a list of candidate single terms ranked by decreasing frequencies, and a list of candidate complex terms ranked by decreasing scores of association strength. The selection of a final set of terms to eventually be acquired requires some threshold tuning, depending on the size of the document collection and the typology and reliability of expected results. Thresholds define *a)* the minimum frequency for a candidate term to enter the lexicon, and *b)* the overall percentage of terms that are promoted from the ranked lists.

3.2. Term organization and structuring

In the second extraction step, proto–conceptual structures involving acquired terms are identified. The basic source of information is no longer a chunked text, but rather a dependency–annotated text, including information about multi–word terms acquired at the previous extraction stage.

We envisage two levels of conceptual organization. Terms in the TermBank are first organized into fragments of head–sharing taxonomical chains, whereby *commercio dei medicinali* 'trade of medicines' and *commercio elettronico* 'elettronic trade' are classified as co–hyponyms of the general single term *commercio* 'trade'. In this way, single and multi–word terms are structured in vertical relationships providing fragments of taxonomical chains.

The second structuring step consists in the identification of clusters of semantically related terms, carried out on the basis of distributionally–based similarity measures. This involves use of CLASS, a distributionally–based algorithm for building classes of semantically–related terms [1]. According to CLASS, two terms are semantically related if they can be used interchangeably in a statistically significant number of syntactic contexts. For all terms (both single and complex) in the TermBank, we extracted from the dependency–annotated text all relations involving these terms in the text. For each term, we selected all the grammatical dependencies it is involved in, and identified (after discarding auxiliary and commonest verbs) the most meaningful (i.e. selective) verbs as resulting from the log–likelihood ratio association measure. The cluster of terms semantically related to a given term is finally determined by computing all the similar terms with respect to each meaningful verb and by grouping the highest ranked terms obtained from the computation on different verbs.

4. Ontology learning from legislative texts: a case study

In this section we summarise the results of a case study carried out on a corpus of legal texts in the environmental domain (Venturi, 2006).

4.1. Corpus description and preprocessing

The corpus consists of 824 legislative, institutional and administrative acts in the environmental domain, for a total of 1.399.617 word tokens, coming from the BGA (*Bollettino Giuridico Ambientale*) database edited by the Piedmont local authority for environment.[1] The corpus includes acts released over a nine years period (from 1997 to 2005) by three different agencies: the European Union, the Italian state and the Piedmont region. It is a heterogeneous document collection including legal acts such as national and regional laws, european directives, legislative decrees as well as administrative acts such as ministerial circulars, decisions, etc.

4.2. The legal–environmental TermBank

Table 1 contains a fragment of the automatically acquired TermBank. For each selected term, the table reports its prototypical form (in the column headed "Term") and its frequency of occurrence in the whole document collection. The choice of representing a domain term through its prototypical form rather than the lemma exponent follows from the assumption that a bootstrapped glossary should reflect the actual usage of terms in texts. In fact, domain-specific meanings are often associated with a particular morphological form of a given term (e.g. the plural form). This is well exemplified in Table 1 where the acquired terms headed by *acqua* 'water' can be parted into two groups according to their prototypical form: either singular (e.g. *acqua potabile* 'drinkable water') or plural (e.g. *acque superficiali* 'surface runoff'). Note, however, that reported frequencies are not limited to the prototypical form, but refer to all occurrences of the abstract term.

ID	Term	Freq
2192	acqua calda	11
974	acqua potabile	36
501	acqua pubblica	121
47	acque	1655
2280	acque costiere	10
2891	acque di lavaggio	6
2648	acque di prima pioggia	8
3479	acque di transizione	5
1984	acque meteoriche	12
1690	acque minerali	16
400	acque reflue	231
505	acque sotterranee	120
486	acque superficiali	131
2692	acque utilizzate	8

Table 1: A fragment of the automatically acquired TermBank

Most notaby, the acquired TermBank includes both legal and environmental terms. The two classes of terms show quite different frequency distributions and turn out to be differentially sensitive to varying frequency thresholds (see Section 3.1). Evaluation of acquired results was carried out with respect to the most conservative TermBank of 4.685 terms, obtained by setting a high minimum frequency threshold (7). Due to the heterogeneous nature of acquired terms, belonging to both the legal–administrative and environmental domains, different resources were taken as evaluation standards: the *Dizionario giuridico* (Edizioni Simone) available online[2] was used as a reference resource for what concerns the legal domain (henceforth referred to as Legal_RR), and the *Glossary of the Osservatorio Nazionale sui Rifiuti* (Ministero dell'Ambiente) available online[3] for the environmental domain (henceforth referred to as Env_RR), which contain respectively 6.041 and 1.090 terminological entries recorded in their prototypical form. For evaluation purposes, more charitable matching metrics between acquired and target terms were considered than full matching, namely:

1. the acquired and target term can appear in different prototypical forms (e.g. *accordi di programma* 'programmatic agreement/PLUR' vs. *accordo di programma* 'programmatic agreement/SING', or *acquisizione dati* 'data acquisition' vs. *acquisizione* ***di*** *dati* 'acquisition of data');
2. the target term can be more general than the acquired one: for example the T2K term *abrogazione di norme* 'repeal of rules' is a good match of Legal_RR *abrogazione* 'repeal';
3. the target term can be more specific than the acquired one: e.g. T2K *agente di polizia* 'policeman' (T2K) is matched against *agente di polizia giudiziaria* 'prison guard' attested in Legal_RR.

Finally, criteria 2 and 3 above can combine with 1. Results are summarised as follows: we found 51% of either full

[1] http://extranet.regione.piemonte.it/ambiente/bga/

[2] http://www.simone.it/cgi-local/Dizionari/newdiz.cgi?index,5,A

[3] http://www.osservatorionazionalerifiuti.it/ShowGlossario.asp?L=Z

or partial matches between the T2K glossary and the reference resources. 89% of the matches covered legal terms and 34,5% environmental ones. 23,5% were found to match entries in both legal and environmental resources. What about the remaining 49% mismatches? How many of them can be considered out–of–dictionary hits? To answer these questions, we selected two additional terminological resources available on the Web: the list of keywords used for the online query of the *Archivio DoGi (Dottrina Giuridica)*[4] for the legal domain, and the thesaurus *EARTh (Environmental Applications Reference Thesaurus)*[5] for the environmental domain. Results are quite encouraging: inclusion of these richer reference resources increases the percentage of matches up to 75,4%. The same percentage goes up even further (83,7%) if we include terms which, in spite of their absence in the selected reference resources, were manually evaluated as domain–relevant terms (see e.g. *anidride carbonica* 'carbon dioxide' in the environmental domain or *beneficiari* 'beneficiary' in the legal one).

5. Conclusions and further directions of research

We reported encouraging results of the application of an automatic ontology learning system, T2K, on a corpus of Italian legislative texts in the environmental domain. Our work shows that the incremental interleaving of robust NLP and machine–learning technologies is key to any attempt to successfully face what we termed the acquisition paradox. By bootstrapping base domain–specific knowledge from texts through knowledge–poor language tools we can incrementally develop more and more sophisticated levels of content representation. In the end the purported dividing line between language–knowledge and domain–specific knowledge proves to be untenable in language use, where language structures and bits of world–knowledge are inextricably intertwined.

There is an enormous potential for this bootstrapping technology. Acquired TermBanks can be transformed into semantic networks linking identified legal and environmental entities. Current lines of research in this direction include a) semi–automatic induction and labelling of ontological classes from the proto–conceptual structures identified by T2K, and b) the extension of the acquired ontology with concept–linking relations (Venturi, 2006). Furthermore, estabilishing the domain relevance of each acquired term represents a central issue in dealing with domain–specificity. By comparing TermBanks automatically extracted from different legislative corpora, we can be successful in classifying the terms belonging to their intersection as specific of the shared domain (in line with the contrastive approach to term extraction proposed by [5]).

6. References

[1] Allegrini, P., Montemagni, S. and V. Pirrelli. Example-Based Automatic Induction Of Semantic Classes Through Entropic Scores. *Linguistica Computazionale*, 1-43: 2003.

[2] Bartolini, R., Lenci, A., Montemagni, S. and V. Pirrelli. Grammar and Lexicon in the Robust Parsing of Italian. Towards a Non-Nave Interplay. In *Proceedings of the International COLING-2002 Workshop "Grammar Engineering and Evaluation"*, Taiwan 2004.

[3] Bartolini, R., Lenci, A., Montemagni, S. and V. Pirrelli. Hybrid Constrains for Robust Parsing: First Experiments and Evaluation. In *Proceedings of LREC 2004*, Lisbon 2004.

[4] Bartolini, R., Giorgetti, D., Lenci, A., Montemagni, S. and V. Pirrelli. Automatic Incremental Term Acquisition from Domain Corpora. In *Proceedings of the 7th International conference on "Terminology and Knowledge Engineering" (TKE2005)*, Copenhagen 2005.

[5] Basili, R., Moschitti, A., Pazienza, M.T. and Zanzotto, F.M. A contrastive approach to term extraction. In *Proceedings of the 4th Conference on Terminology and Artificial Intelligence (TIA2001)*, Nancy, France, 2001.

[6] Buitelaar, P., Cimiano, P., and B. Magnini. Ontology Learning from Text: an Overview. In Buitelaar et al. (eds.), *Ontology Learning from Text: Methods, Evaluation and Applications* (Volume 123 Frontiers in Artificial Intelligence and Applications): 3–12, 2005.

[7] Dell'Orletta, F., Lenci, A., Marchi, S., Montemagni, S. and V. Pirrelli. Text-2-Knowledge: una piattaforma linguistico-computazionale per l'estrazione di conoscenza da testi. In *Proceedings of the SLI-2006 Conference*: 20–28, Vercelli 2006.

[8] Dunning, T. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*: 19(1), 1993.

[9] Federici, S., Montemagni, S. and V. Pirrelli. Shallow Parsing and Text Chunking: a View on Underspecification in Syntax. In *Proceedings of the Workshop On Robust Parsing*, in the framework of the European Summer School on Language, Logic and Information (ESSLLI-96), Prague 1996.

[10] Lame, G. Using NLP techniques to identify legal ontology components: concepts and relations. *Lecture Notes in Computer Science*, Volume 3369: 169–184, 2005.

[11] Sais, J. and P. Quaresma. A Methodology to Create Legal Ontologies in a Logic Programming Based Web Information Retrieval System. *Lecture Notes in Computer Science*, Volume 3369: 185–200, 2005.

[12] Valente, A. Types and Roles of Legal Ontologies. *Lecture Notes in Computer Science*, Volume 3369: 65–76, 2005.

[13] Venturi, G. L'ambiente, le norme, il computer. Studio linguistico–computazionale per la creazione di ontologie giuridiche in materia ambientale. Degree Thesis, Manuscript, December 2006.

[14] Walter, S. and M. Pinkal. Automatic extraction of definitions from german court decisions. In *Proceedings of the COLING-2006 Workshop on Information Extraction Beyond The Document*: 20–28, Sidney 2006.

[4] http://nir.ittig.cnr.it/dogiswish/dogiConsultazioneClassificazioneKWOC.php

[5] http://uta.iia.cnr.it/earth.htm#EARTh%202002