NLP-based ontology learning from legal texts. A case study.

Alessandro Lenci[1], Simonetta Montemagni[2], Vito Pirrelli[2], Giulia Venturi[2]
[1] *Dipartimento di Linguistica Ũ Università di Pisa, Italy*
[2] *Istituto di Linguistica Computazionale - CNR, Italy*

Abstract. The paper reports on the 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.

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

1. Introduction

Ontology building is nowadays a very active research field, as witnessed by the fast growing literature on the topic and the increasing number of Knowledge Management applications based on automated routines for ontology navigation and update. The enterprise, however, requires harvesting domain–specific knowledge on an unprecedented scale, by tapping and harmonizing knowledge sources of highly heterogeneous conception, format and coverage, ranging from foundational ontologies and structured databases to electronic text documents. As electronic texts still represent the most accessible and natural repositories of specialised information worldwide, we seem to have reached a stage where an unlimited demand for ontologically–interpreted knowledge disproportionally exceeds the availability of automatically–interpreted textual information.

To bridge such a critical gap, different methodologies have been proposed to automatically extract information from texts and provide a structured organisation of extracted knowledge in as diverse domains/sectors as bio–informatics, health–care, public administration and company document bases. The situation in the legal domain is in line with this general trend and probably made even more critical by the fact that laws are invariably conveyed through natural language.

The last few years have seen a growing body of research and practice in constructing legal ontologies and applying them to the law domain. A number of legal ontologies have been proposed in different research projects: yet, most of them focus on a upper level of concepts and were mostly hand-crafted by domain experts (for a survey of legal ontologies, see Valente 2005). It goes without saying that realistically large knowledge-based applications in the legal domain need more comprehensive ontologies incorporating up-to-date knowledge: ontology-learning from texts could be of some help in this direction.

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 Lame (2005), Saias and Quaresma (2005) and Walter and Pinkal (2006). 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 (*Text–to–Knowledge*), 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 (Dell'Orletta *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 personalised knowledge profiling. Prototype versions of T2K are currently running on public administration portals and have been used for indexing E–learning and E–commerce materials. In what follows, we report some ontology learning experiments carried out with T2K on Italian legislative texts.

2. From text to knowledge: the role of NLP tools

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 ambigui-

ties, 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 perspectivised and combinatorially ambiguous information bits.

In our view, there is no simple way out of this paradox. Pattern matching techniques allow for fragments of knowledge to be tracked down only in very limited text windows, while foundational ontologies are too general to be able to make successful contact with language variability at large. The only effective solution, we believe, is to understand and 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 syntagmatic units and organizing them into non-recursive phrasal "chunks" that do not seem to require domain-specific knwoledge. In turn, chunked texts can sensibly be accessed and compared for statistically–significant 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 them in 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 pos–tagged 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 differentially 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", see below). Identification of conceptual structures, on the other hand, is carried out against a dependency–annotated text. In what follows, the general architecture of the Italian parsing system underlying T2K (henceforth referred to as AnIta, Bartolini *et al.*, 2004) is briefly illustrated.

2.1. An outline of AnIta

The AnIta system consists of a suite of linguistic tools in charge of:

1. tokenisation of the input text;
2. morphological analysis (including lemmatisation) of the text;
3. parsing, articulated in two different steps:
   a) "chunking", carried out simultaneously with morpho-syntactic disambiguation;
   b) dependency analysis.

In what follows we will focus on the syntactic parsing components in charge of the linguistic pre-processing of texts for the different ontology learning tasks of T2K.

Text chunking is carried out through a battery of finite state automata (CHUG–IT, Federici *et al.*, 1996), which takes as input a morphologically analysed and lemmatised text and segments it into an unstructured (non-recursive) sequence of syntactically organized text units called "chunks" (e.g. nominal, verbal, prepositional chunks). Chunking requires a minimum of linguistic knowledge; its lexicon contains no other information than the entry's lemma, part of speech and morpho–syntactic features. A chunk is a textual unit of adjacent word tokens sharing the property of being related through dependency relations (es. pre–modifier, auxiliary, determiner, etc.). A chunked sentence, however, does not give information about the nature and scope of inter–chunk dependencies which are identified during the phase of dependency analysis (see below). Morpho–syntactic disambiguation is performed simultaneously to the chunking process.

Il presente decreto stabilisce le norme per la prevenzione ed il contenimento dell'inquinamento da rumore [...]
'this decree establishes the rules for prevention and control of noise pollution [...]'

```
[[CC:N_C][DET:IL#RD@MS][PREMOD:PRESENTE#A@MS][POTGOV:DECRETO#S@MS]]
[[CC:FV_C][POTGOV:STABILIRE#V@S3IP]]
[[CC:N_C][DET:LO#RD@FP][POTGOV:NORMA#S@FP]]
[[CC:P_C][PREP:PER#E][DET:LO#RD@FS][POTGOV:PREVENZIONE#S@FS]]
[[CC:COORD_C][CONJTYPE:E#CC]]
[[CC:N_C][DET:IL#RD@MS][POTGOV:CONTENIMENTO#S@MS]]
[[CC:di_C][DET:LO#RD@MS][POTGOV:INQUINAMENTO#S@MS]]
[[CC:P_C][PREP:DA#E][POTGOV:RUMORE#S@MS]]
```

Figure 1. A sample of chunked text

To be more concrete, the sentence fragment reported in Figure 1 is segmented into eight chunks, each including a sequence of adjacent word tokens mutually related through dependency links of some kind. For example, the first nominal chunk (N_C) covers three word tokens, *il presente decreto*: the noun head *decreto*, the adjectival premodifier *presente* and an introducing definite article. Although the representation is silent about the relationship between *stabilire* 'establish' and *le norme* 'the rules', this is not to entail that such a relationship cannot possibly hold: simply, the lexical knowledge available to this parsing component makes it impossible to state unambiguously how chunks relate to each other and the nature of this relationship. This is the task for further analysis steps.

Dependency parsing is aimed at identifying the full range of syntactic relations (e.g. subject, object, modifier, complement, etc.) within each sentence: syntactic relations are represented as dependency pairs between lexical heads. It is carried out by IDEAL (Bartolini *et al.*, 2002), a finite state compiler for dependency grammars. The IDEAL general grammar of Italian is formed by ca. 100 rules covering the major syntactic phenomena. The grammar rules are regular expressions (implemented as finite state automata) defined over chunk sequences, augmented with tests on chunk and lexical attributes. A "confidence value" (PLAUS) is associated with identified dependency relations, to determine a plausibility ranking among competing analyses. Figure 2 reports the dependency representation of the same sentence.

The output consists of binary relations between content words, typically a head and a dependent. There may be features associated with both participants in the relation conveying other types of information such as the semantic type of a dependent (ROLE) or the preposition

```
MODIF(DECRETO[34544.1],PRESENTE[34544.1]<role=RESTR>)plaus=100
SUBJ(STABILIRE[34544.2],DECRETO[34544.1])plaus=50
OBJD(STABILIRE[34544.2],NORMA[34544.3])plaus=50
COMP(NORMA[34544.3],PREVENZIONE[34544.4]<intro=PER>)plaus=50
COORD(PREVENZIONE[34544.4],CONTENIMENTO[34544.6]<role=CONJ>)plaus=50
ARG(CONTENIMENTO[34544.6],INQUINAMENTO[34544.7]<intro=DI>)plaus=60
COMP(INQUINAMENTO[34544.7],RUMORE[34544.8]<intro=DA>)plaus=50
```

Figure 2. A sample of dependency–parsed text

introducing a certain relation (INTRO). The sentence fragment is described by 7 dependency relations including subject, object as well as other modification relations: for instance, *decreto* has been identified as the subject of the verb *stabilire* and *norme* as its direct object.

There are some reasons to believe that chunked texts are a suitable starting point for term extraction from a continuously expanding document base. First, thanks to its knowledge–poor lexicon, chunking is fairly domain–independent. Moreover, its finite–state technology makes chunking very robust and flexible in the face of parse failures: unparsed sequences are tagged as unknown chunks and parsing can resume from the first ensuing word–form which is part of a parsable chunk. Thirdly, chunking provides a first level of syntactic grouping which, however crude, paves the way to reliable and wide–coverage identification of candidate domain terminology, including both single and multi–word terms. As chunks standardise a considerable amount of grammatical information, searching for candidate terms in a chunked text can be done at a considerable level of abstraction from language nitty–gritty. On the other hand, identification of clusters of semantically related terms or acquisition of relations between terms constitute more demanding tasks requiring deeper levels of linguistic analysis such as dependency parsing.

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 a) fragments of taxonomical chains, and b) clusters of semantically related terms.

Figure 3 illustrates the functional architecture of T2K:

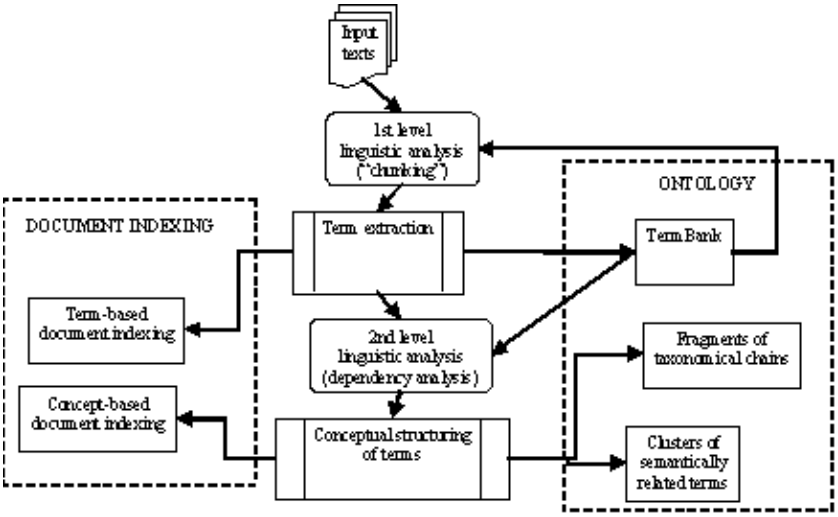

Figure 3. T2K architecture

The two basic steps take the central pillar of the portrayed architecture, showing the interleaving of 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 (Buitelaar *et al.*, 2005), 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. Terms are surface realisations of domain–specific concepts and represent, for this reason, a basic prerequisite for more advanced ontology learning tasks. In principle, they need be recognized whatever the surface form they show in context, irrespectively of morpho–syntactic and syntactic variants. 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). Term extraction thus requires some level of linguistic pre-processing of texts.

T2K looks for terms in syntactically chunked texts such as those illustrated in Section 2.1 (Figure 1). 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 chunked source 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, interpreted and applied by the IDEAL compiler. The set of chunk patterns covers the main types of modification observed in complex nominal terms: i.e. adjectival modification (e.g. *organizzazione internazionale* 'international organisation'), prepositional modification (e.g. *tutela del territorio* 'protection of the territory'), including more complex cases where different modification types are compounded (e.g. *incenerimento dei rifiuti pericolosi* 'incineration of dangerous waste'). Secondly, the list of acquired potential complex terms is ranked according to their log–likelihood ratio (Dunning, 1993), 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. We tested the log–likelihood ratio against other association measures such as mutual information, chi–square etc., log-likelihood faring consistently better than the others. Moreover this measure is known to be less prone to assigning high scores to very sparse pairs. It should be recalled that the log–likelihood ratio is commonly used for discovering collocations. Hence, we are treating complex terms as though they belonged to the more general class of collocations. However, T2K uses the log–likelihood ratio in a somewhat atypical way: instead of measuring the association strength between adjacent words, T2K measures it between the lexico–semantic heads of adjacent chunks. The main and often underestimated advantage of defining co–occurrence patterns over syntactic structures is that we can broaden our search space (the text window) in a controlled way, by making it sure that there is a syntactic pattern linking two adjacent lexical heads.

So far, acquisition of potential complex terms has involved chunk pairs only (bigrams). In T2K recognition of longer terms is carried out by iterating the extraction process on 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 (Bartolini *et al.*, 2005). 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 only at later stages. This is the case, for instance, of coordination patterns.

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 be included in the TermBank 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. Typical values for a corpus of about one million tokens are as follows: minimum frequency threshold equal to 7 for both single and complex terms; selected single terms are the topmost 10% in the ranked list; selected multi-word terms are the topmost 70% in the ranked list of potential complex terms.

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 the dependency–based analysis exemplified in Figure 2, with the original text containing an explicit indication of the multi–word terminology 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 *ambiente urbano* 'urban environment' and *ambiente marino* 'marine environment' are classified as co–hyponyms of the general single term *ambiente* 'environment'.

Moreover, T2K clusters semantically–related terms by using CLASS, a distributionally–based algorithm for building lexico–semantic classes (Allegrini *et al.*, 2003). According to CLASS, two terms are semantically related if they can be used interchangeably in a statistically significant number of syntactic contexts. The starting point for the CLASS algorithm is provided by a dataset of dependency triples – $< T, C, s >$ –, where T is a target linguistic expression, C is a linguistic context for T, and s is the particular syntagmatic dependency relation between T and C. For our present concerns, variables are interpreted as follows:

1. T corresponds to an acquired term in the TermBank;

2. s stands for either a subject or a direct object dependency relation;

3. C corresponds to a verb with which T is attested to co–occur as a subject or a direct object. In fact, of all verb–term pairs attested in the corpus only a subset of highly salient such pairs is considered for clustering by CLASS. Light verbs such as *take* or *make* are likely to give very little information about the semantic space of the terms they select in context. Hereafter we shall refer to the set of highly salient verbs keeping company with subject/object T as the *best verbs for* T, or BVT. For each term T, BVT contains those verbs only whose strength of association with subject/object T (measured by the log–likelihood ratio) exceeds a fixed threshold.

For all terms (both single and complex) in the TermBank, we extract from the dependency–annotated text the best verb/subject and verb/object pairs. CLASS then computes the degree of semantic relatedness between two terms T_1 and T_2 by measuring the degree of overlapping between BVT_1 and BVT_2, according to the metric described in Allegrini *et al.*, (2003). This corresponds to the assumption that the semantic similarity between two terms is a function of the possibility for the entities denoted by the terms to be involved in similar events, where the latter are expressed by the term best verbs. The cluster of terms semantically related to a target term T is finally ordered by decreasing similarity scores with respect to T. For each term, the user can define the maximum number of related terms to be returned by the system; this parameter can be set on the basis of the user's needs (it should be kept in mind that going down in the ranked list of related terms the semantic distance from T increases; therefore, it becomes more likely to find spurious associations).

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 belonging to the environmental domain (Venturi, 2006).

4.1. Corpus description and preprocessing

The corpus consists of 824 legislative, institutional and administrative acts concerning 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

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

Table I. An excerpt of the automatically acquired TermBank

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

corpus includes acts released by three different agencies, i.e. the European Union, the Italian state and the Piedmont region, which cover a nine years period (from 1997 to 2005). It is a heterogeneous document collection including legal acts such as national and regional laws, european directives, legislative decrees, etc. as well as administrative acts such as ministerial circulars, decisions, etc.

4.2. The legal–environmental TermBank

Table I contains a fragment of the automatically acquired TermBank. For each selected term, the TermBank reports its prototypical form (in the column headed "Term"), its frequency of occurrence in the whole document collection, and the lemma of the lexical head of the chunk covering the term (see column "Lemmatised headwords"). The choice of representing a domain term through its prototypical form rather than the lemma (as typically done in ordinary dictionaries) 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 I 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'). It

should be noted, however, that reported frequencies are not limited to the prototypical form, but refer to all occurrences of the abstract term.

As expected from the peculiar nature of processed documents, the acquired TermBank includes both legal and environmental terms. Since the two classes of terms show quite different frequency distributions, different acquisition experiments were carried out by setting different thresholds (see Section 3.1). By using standard thresholds with respect to corpus size, we obtained a TermBank of 4.685 terms (both single and multi–word terms): the selected minimum frequency threshold for both single and multi-word terms was 7, the percentage of selected terms from the ranked lists was 10% in the case of single terms and 70% for multi–word terms. Yet, in this TermBank, environmental terms were scarcely represented due to their high rank (and low frequency) according Zipf's law. Since the focus of our interest was on both types of terminology, we carried out new acquisition experiments by reducing the minimum frequency thresholds to 5 and 3. In both experiments, the number of acquired environmental terms increased, unfortunately together with noisy terms. For instance, with the minimum frequency threshold set to 3, the numer of extracted terms is more than doubled, i.e. it is equal to 11.103.

Evaluation of acquired results was carried out with respect to the TermBank of 4.685 terms (i.e. the one obtained by setting the minimum frequency threshold equal to 7). Due to the heterogeneous nature of the terms in the glossary, belonging to both the legal–administrative and the environmental domains, two different resources were taken as a gold standard: 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, different types of matches were taken into account. Besides the full match between the T2K term and the term in the reference resource, different types of partial matches were also considered, i.e.:

1. the same term appears both in the T2K TermBank and in the gold standard resource but under different prototypical forms: this is the case, for instance, of the term *accordi di programma* 'programmatic agreement' which appears in the plural form in T2K and in the singular form in Legal_RR. At this level, two terms may also differ

[2] http://www.simone.it/cgi-local/Dizionari/newdiz.cgi?index,5,A
[3] http://www.osservatorionazionalerifiuti.it/ShowGlossario.asp?L=Z

for the prepositions linking the nominal headwords of a complex term, as in the case of *acquisizione dati* vs *acquisizione **di** dati* 'acquisition of data' or *abbandono **di** rifiuti* vs *abbandono **dei** rifiuti* 'waste abandon';

2. the gold reference resource contains a more general term whereas T2K acquired one of its hyponyms: this is the case of the T2K term *abrogazione di norme* 'repeal of rules', which in Legal_RR occurs in its more general form *abrogazione* 'repeal';

3. the reverse case with respect to 2 above, i.e. the gold reference resource contains a more specific term with respect to T2K which extracted a more general term, typically its hyperonym: e.g. *agente di polizia* 'policeman' (T2K) vs *agente di polizia giudiziaria* 'prison guard' (Legal_RR).

In the cases described in 2 and 3 above, a distinction is made – again – between matches concerning the prototypical form and matches at the level of stemmed words.

The results of the evaluation carried out on the basis of the criteria described above can be summarised as follows: in 51% of the cases a match, either full or partial, was found between the T2K glossary and the references resources; in particular, 89% of identified matches was concerned with legal terms and 34,5% with environmental ones, with a 23,5% of terms occurring in both reference resources. The question arising at this point is whether the remaining 49% of terms for which no match was found was represented by errors and noisy terms or were domain–specific terms not appearing in the selected reference resources. In order to answer this question, we selected two additional 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, against which a manual evaluation was carried out for 25% of the automatically acquired T2K glossary. The results are quite encouraging: by including these two richer reference resources, the percentage of matching terms increased to 75,4%. This percentage grows up to 83,7% if we also include terms which, in spite of their absence in the selected reference resources, were manually evaluated as domain–relevant terms: this is the case, for instance, of the terms *anidride carbonica* 'carbon dioxide' for what concerns the environmental domain or *beneficiari* 'beneficiary' for the legal one. The percentage of

[4] http://nir.ittig.cnr.it/dogiswish/dogiConsultazioneClassificazioneKWOC.php
[5] http://uta.iia.cnr.it/earth.htm#EARTh%202002

manually detected errors is 21,1%, also concerning some of the terms for which a partial match was detected. Whereas on the basis of these results it can be claimed that the accuracy of T2K for what concerns term extraction is quite high, nothing can be said as far as recall is concerned. As a matter of facts, the selected reference resources could not be used for this specific purpose due to their wider coverage, not circumscribed to the environmental domain.

4.3. Proto–conceptual organisation of terms

A first step towards the conceptual organization of terms in the TermBank consists in building taxonomical chains. This is to say that single and multi–word terms are structured in vertical relationships providing fragments of taxonomical chains such as the one reported below:

```
applicazione
   applicazione dei paragrafi
   applicazione dell' articolo
   applicazione della direttiva
   applicazione della legge
   applicazione della tariffa
   applicazione delle disposizioni
   applicazione delle sanzioni
      applicazione delle sanzioni amministrative
      applicazione delle sanzioni previste
   applicazione del presente decreto
   applicazione del regolamento
   applicazioni di quarantena
```

where the acquired direct and indirect hyponyms of the term *applicazione* 'enforcement' are reported. In this example, it can be noticed that terms sharing the head only are the direct hyponyms of the root term. Further hyponymy levels can be detected when two or more multi–word terms share not only the head but also modifiers, as in the case of the *applicazione delle sanzioni amministrative* 'enforcement of administrative sanctions' with respect to the more general term *applicazione delle sanzioni* 'enforcement of sanctions'.

With minimum frequency threshold set to 7, the numer of extracted hyponymic relations is 2.181 referring to 272 hyperonym terms; with the threshold set to 3, identified hyponymic relations increase to 6.635 regarding 454 hyperonym terms.

The second structuring step performed by T2K consists in the identification of clusters of semantically related terms which is carried out on

the basis of distributionally–based similarity measures (see Section 3.2). In what follows, clusters of semantically related terms are exemplified for both domains:

```
disposizioni 'provision'
   norme, disposizioni relative, decisione, atto, prescrizioni
legge 'law'
   regolamento, protocollo, accordo, statuto, amministrazioni comunali
inquinamento 'pollution'
   danno ambientale, inquinamento marino, effetti nocivi, conseguenza,
   inquinamento atmosferico
impatto ambientale 'environmental impact'
   esposizione, danno, esigenze, conseguenza, pericolo
```

For each target term, the set of the first 5 most similar terms is returned, ranked for decreasing values of semantic similarity. With the minimum frequency threshold set to 7, the number of identified related terms is 3.448 referring to 665 terminological headwords.

As illustrated in Section 3.2, these clusters of related terms were computed with respect to the most salient verbs associated with each target term: for instance, for *disposizione* 'provision' the most strongly associated verbs included *applicare* 'enforce', *adottare* 'pass', *abrogare* 'repeal', *decorrere* 'to have effect from' etc., whereas for *inquinamento* 'pollution' they range from *combattere* 'fight against', *ridurre* 'reduce', *prevenire* 'prevent', *eliminare* 'eliminate' to *causare* 'cause', *provocare* 'bring about' and *controllare* 'watch'. The terms similarity chains resulting from context-sensitive similarity measures are then merged and ranked according to decreasing similarity weights. It should be appreciated that in these clusters of semantically related words different classificatory dimensions are inevitably collapsed; they include not only quasi–synonyms (as in the case of *disposizioni* 'provision' and *norme* 'regulations' or *inquinamento* 'pollution' and *danno ambientale* 'environmental damage'), hyperonyms and hyponyms (e.g. *inquinamento* 'pollution' and *inquinamento atmosferico* 'atmospheric pollution'), but also looser word associations. As an example of the latter we mention the relation holding between *legge* 'law' and *amministrazione comunale* 'municipal administration', or between *pericolo* 'danger' and *conseguenza* 'consequences' and the environmental term *impatto ambientale* 'environmental impact'.

5. Conclusions and further directions of research

We reported preliminary but extremely 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 the 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 (first steps in this direction are reported in Venturi, 2006).

Our experiments also hightlighted some interesting open issues which need to be tackled in the near future. As pointed out in Section 4.2, running T2K on a corpus of legislative and administrative acts results in a two–faced terminological glossary, which includes terms belonging to both the legal–administrative and environmental domains. Estabilishing the domain relevance of each acquired term represents a central issue when dealing with legal–administrative texts. Some preliminary experiments have already been carried out in order to semi–automatically identify the domain–relevance of each acquired term. In particular, terminology acquisition was carried out with T2K on thematically different legislative corpora. By comparing the TermBanks automatically extracted from different corpora, we could classify the terms belonging to their intersection as belonging to the legal–administrative lexicon. This is in line with the contrastive approach to term extraction proposed by Basili *et al.* (2001). Similarly, the relevance of environmental terms will be validated by running terminology extraction on the environmental literature.

References

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

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

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.

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.

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.

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.

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.

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

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.

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

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.

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

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.

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.