

Language Technologies Meet Ontology Acquisition

Galia Angelova

Institute for Parallel Processing, Bulgarian Academy of Sciences
25A, Acad. G. Bonchev Str., 1113 Sofia, Bulgaria, galia@lml.bas.bg

Abstract. This paper overviews and analyses the on-going research attempts to apply language technologies to automatic ontology acquisition. At first glance there are many successful approaches in this very hot field. However, most of them aim at the extraction of named entities as well as draft taxonomies and paronomies. Only few attempts exist for enriching ontologies by applying word-sense disambiguation. There are principle obstacles to extract automatically coherent conceptualisations from raw texts: it is impossible to identify exactly the types and their instances as well as the word meanings which denote types. It is also impossible to validate a text-based conceptual model against the real world. Thus we can expect only partial success in the semi-automatic acquisition in specific (limited) domains, by workbenches supporting the human knowledge engineer in the final ontological choices.

Keywords: natural language processing, information extraction, automatic knowledge acquisition from text

1 Introduction

Recent developments in artificial intelligence, knowledge representation, WWW, and information-processing applications resulted in the advent of the Semantic Web [1]. Its ultimate aim is to make the web resources more meaningful to computers by augmenting the presentation markup with semantic markup, i.e. meta-data annotations that describe the content. It is widely expected that the innovation will be provided by agents and applications dealing with ontology acquisition, merging and alignment, annotation of www-pages towards the underlying ontologies as well as intelligent, semantic-based text search and intuitive visualisation. However, the current progress in all these directions is not very encouraging despite the impressive number of running activities. Isolated results and tools are available for e.g. automatic and semi-automatic annotation of web pages, for knowledge-based information retrieval, for partial ontology learning and so on but it is still difficult to grasp a coherent picture of how the Semantic Web will drastically change the information age by offering quite new kinds of services. Another discouraging obstacle is that the ontologies for the Semantic Web are not clearly seen at the horizon. And, after five years of investments and active investigation, the vision of the “universal” Semantic Web becomes a “futuristic” target. Instead, the goals shift to realistic developments and notions like “*intelligent semantic-based Web applications*”, “*ontology-based tools*”, “*application ontologies*”, “*light-weight ontologies*”, “*bridging corporative views via negotiation*”,

2 Galia Angelova

“mediation between entities in order to integrate them”, “use cases which demonstrate the viability of the approach”¹ and so on.

In this way the development of the Semantic Web and the ontology-driven applications is currently slowed down due to principal problems, among them the knowledge acquisition bottleneck. Therefore, language technologies for free text processing are extensively applied to create or grow ontologies “in a period as limited as possible with a quality as high as possible” (citation from [2], a nice summary of the dominating project-oriented perspective). We can certainly remind that the idea of automatic knowledge acquisition from text is an older dream. However, the present achievements rely on (i) advanced language processing tools and (ii) very large linguistics resources – two artifacts that were not available twenty or thirty years ago. That is why the activities in this sphere deserve careful observation and analysis, as they are attempts to process all kinds of raw texts and really massive amounts of data.

This article analyses the on-going work in automatic knowledge acquisition by choosing several representative papers (instead of listing hundreds of titles), which contain research and application innovation from a language technology perspective. In addition, they try to map the results (extracted language units) to ontological elements and structures. We consider the latter mapping a decisive step in the automatic knowledge acquisition. Due to this reason we skip here many papers which deal with natural language understanding and word-sense disambiguation from the computational linguistics perspective, as they are focused on extraction of text semantics rather than on world knowledge. Section 2 overviews the progress in Named Entity Recognition (NER) – which is in fact a collection of NE instances in texts with annotations of their types. Section 3 presents approaches for learning of concept types from texts, by clustering of natural language terms into semantic sets standing for domain concepts. Section 4 summarises the achievements in detection of conceptual relationships and automatic constructions of taxonomies and enlargement of ontologies. Section 5 discusses kinds of ontological constructions which are *not* targets of automatic knowledge acquisition today.

2 Acquisition of Instances (Named Entities)

The tasks of identification, extraction and classification of named entities from texts have attracted special attention in the 80-ies (last century) in the emerging area of Information Extraction (IE) and Message Understanding [3]. Indeed, all named entities – proper names, names of organizations, companies, locations, even dates – bear much information about the text content, genre, sometimes author etc. NER is usually approached by grammar-based text analysis. Often the extracted text items are mapped to very large lexicons of names which are created in advance. The collection of such multilingual lexicons is an important IE activity. The IE systems are compared at special competitions according to several parameters, among them the NER success rate. NER is evaluated via the percentage of correctly recognized named entities which is a measure for the grammar coverage as well as an indirect hint for

¹ Citations from the Semantic Web session at the event “Information Society Technologies 2004” (IST-04), organised by the European Commission, The Hague, 15-17 November 2004.

Entity	Algorithm
Abstract	Application
Happening	Application_Domain
Object	Event
Agent	Conference
Organisation	Tutorial
Person	Workshop
BusinessObject	Formal_Language
InformationResource	Method
Location	Organisation
AstronomicalObject	Association
Facility	Department
GlobalRegion	Enterprise
LandRegion	Institute
NonGeographicalLocation	Research_Funding_Institution
PoliticalRegion	University
PopulatedPlace	Person
StreetAddress	Project
WaterRegion	Publication
Product	Book
Statement	InProceedings
Vehicle	Report
	Tool
	Topic

Fig.1. *Left*: KIM Ontology. *Right*: a sample ontology in OntoMat Annotizer

the amount of the available lexical resources of names. The most advanced IE systems today recognize some kinds of named entities in unknown texts with more than 90% accuracy. For instance, KIM [4] recognizes English *person names* and *locations* in inter-domain web content with 89.09% and 91.23% accuracy correspondingly². KIM's current lexicon of names (resource of instances) contains more than 200,000 items. Each location has several aliases (English, French, Spanish and sometimes the local transcription). The IE systems need such data, because locations are difficult to recognize otherwise. KIM tries to extract more information than the isolated instances; it searches for patterns defining attributes and relations of the featured entity, like: *subRegionOf* property for *Location-s*, *hasPosition* for *Persons*, *locatedIn* for *Organizations*, etc. The automatic extraction of such features is much less successful than the recognition of single, isolated named entities.

The essence of named entities extraction is to recognize them as instances according to some ontology of concepts. Figure 1 presents the KIM ontology, which is used for automatic classification of names. This ontology consists of 250 classes and has about 100 relations. Ontologies are also applied for manual annotation – i.e. classification - of named entities. A sample ontology in OntoMat Annotizer [5] is

² More precisely, IE applies the classical *recall* R (the ratio of correctly extracted entities against all the available entities present in the texts) and *precision* P (the ratio of correctly extracted entities against all the extracted entities). What we called accuracy above is the harmonic mean of R and P – known as *F-measure*, where $F = (2 * P * R) / (P + R)$.

given at Figure 1 too. The two ontologies clearly illustrate that there are many kinds of named entities in the texts and obviously, their automatic recognition and classification require substantial efforts. The NER result - extracted “drafts of named entities” - needs further manual validation and human correction, in case we want to build correct lists of concept instances with some associated attributes.

In the context of NER, let us comment on several issues which are relevant to our discussion regarding automatic ontology acquisition.

First, we emphasize that the extracted named units are symbol strings that are “mechanistically” collected from different texts. For instance, the NER module might extract *Sofia* as a person name (most probably female) and as a city name (hopefully a city located in Bulgaria). But it is unlikely to extract information that there was a Byzantine princess *Sofia* who lived in the city of *Sofia*, at present in Bulgaria. This fact is too complex to be extracted at the NER stage. Usually only titles and professional positions are recognized successfully “around” the person names (e.g. *President Bush*). In addition, the task of NER cannot infer that many women are named *Sofia* or *Sophia* and often the two names are language-dependent variants. So the result of NER is a list of isolated items tagged via ontology types.

Second, the NER task does not distinguish well between different variants of the same name. For instance, “*The Bulgarian Academy of Sciences*” is organization name which can appear in the text as “*The Academy of Sciences*” or simply “*The Academy*” in some unambiguous contexts (while *the academy* often refers to the main building of the organization). In addition NER does not resolve the references in the text as this requires deep natural language understanding. However, the recognition of referential citations is important prerequisite for successful acquisitions of facts. Consider the following discourse, which consists of sentence 1 and sentence 2:

Sofia was a Byzantine princess. She lived in the city of Serdica, known today as Sofia.

The resolution of the pronominal anaphor “*she*” in sentence 2 to “*Sofia*” in sentence 1 is a necessary condition for extraction of the historical fact we want to encode. Now we make the following important observation: even the simplest facts regarding instances might be communicated by complex language structures in free texts. One may assume that the recognition of named entities is simpler – which is true in a sense – but the extraction of related facts may require implementation of the full potential of natural language processing and natural language understanding, which looks impossible today. The variety of natural language expressions is so high that important facts cannot be encoded without ambiguity even in controlled languages (at least there are still no experimental evidences for optimism beyond the naturally restricted languages of some very specific technical domains, e.g. aircraft industry).

So at this point of our discussion we already realise why we lack successful applications which acquire knowledge about arbitrary entities from free texts. Today propositions are acquired only from suitably formulated text statements. Due to the “natural understanding bottleneck”, the automatic ontology construction is (still) not an alternative which may replace the precise manual definition and construction. The present approaches are focused on the acquisition of domain concepts as well as relevant statements that are easy to identify and extract. Another important activity is the integration of concepts into ontologies.

3 Acquisition of Concepts from Texts

The main machine learning approach, applied to extract concepts (semi-) automatically from texts, is clustering of co-occurrences of words. The idea is that similar words appear as collocations with the same verbs. Moreover, words are similar to the extent they appear in similar contexts (this is the so called distributional hypothesis). No annotation of the input text is needed beforehand. Usually the first step is to perform lexical and morphological analysis of the input. Then complex terms of several tokens are grouped as single units and afterwards (partial) syntactic analysis is performed to extract main sentence phrases or to fully parse each sentence. Below we briefly summarize some relevant activities and their results, keeping our focus to the language technologies involved in the process.

The paper [6] presents techniques applied in the ontology learning environment *Text-To-Onto*, such as ontology learning from free text, from dictionaries, or from legacy ontologies. A convincing example illustrates the benefit of processing a domain thesaurus, where the concepts are described together with their definitions and hierarchy. The environment *Text-To-Onto* was applied to a machine-readable dictionary of an insurance company which contained entries like the following one:

***Automatic Debit Transfer:** Electronic service arising from a debit authorization of the Yellow Account holder for a recipient to debit bills that fall due direct from the account.*

Several heuristics are applied to this morphologically analysed definition. One simple heuristic relates the definition term *automatic debit transfer* to the first noun phrase occurring in the definition - *electronic service*. Their corresponding concepts are linked in a draft hierarchy:

AUTOMATIC DEBIT TRANSFER IS-A ELECTRONIC SERVICE.

Applying this heuristic iteratively, one may propose large parts of the target ontology. The process of ontology learning is semi-automatic with human intervention, as the obtained hierarchy is further refined by a human engineer using the ontology engineering workbench *OntoEdit*. In this way, specific language resources like thesauri are very useful for the concept acquisition phase. However, often dictionaries with definitions are not available and then the acquisition starts from free texts.

The papers [2,7] present systematic attempts to build an ontology from scratch. The research is done within the project *OntoBasis*, which deals with elaboration and adaptation of text analysis tools for the construction of specific domain ontologies. Here we use a very simple example to illustrate the approach. The tools extract automatically binary relations (lexons) by mining the Verb-Object dependency, applying selectional restrictions and functional relations. In fact they extract pairs

[*Main Verb - Nominal String*],

where the Nominal String is a string of adjectives and nouns. Then the nominal strings are clustered according to the cooccurring verbs. A sample cluster from a Hepatitis corpus is:

6 Galia Angelova

{ *liver transplantation, transplplantation, orthotopic liver transplantation* }.

The suggested concept is described in the corresponding medical text by three nominal strings.

In addition to the mining of the verb-object pairs, pattern matching is done for triples

[*Nominal string - Preposition - Nominal string*].

The resulting phrases look as follows:

```
blood_vessel_growth on ribonucleolytic_activity,  
amino_acid_residue in polymerase,  
primer from amino_acid_sequence.
```

These triples are organised in classes of prepositional structures and compared to the clusters obtained after mining the verb-object pairs. In case of similarity, clusters are augmented to the following extracts:

[*Nominal String Preposition AugmentedCluster*].

The obtained “correct” structures look as follows:

```
[dose, injection, vaccination] of  
[hepatitis B vaccine, HBV vaccine, vaccine]  
[use] of [face mask, mask, glove, protective eyewear]  
[vaccination, vaccine] against [disease, virus, virus type]
```

Other resulting constructs are:

```
[level, expression] of ...  
[effect] on ...  
[increase] in ...
```

Obviously, these lists need further human intervention for refinement and organization into ontology.

This unsupervised approach for ontology initiation looks promising, as it requires no preliminary tagging of the training data and relies on relatively simple language technologies and tools which are more or less available for many natural languages. However, it is not very successful. It was evaluated against UMLS (the Unified Medical Language System) by comparison of the extracted nominal strings to the UMLS labels. The *recall* (percentage of correctly recognized nominal strings against all relevant nominal strings) and the *precision* (percentage of correctly recognized nominal strings against all recognized nominal strings) were computed according to the quantity of UMLS pairs found in the clusters. The reported results for the Hepatitis corpus are less than 33% recall and less than 17% precision, for clustering of 100-500 words [7]. Adding prepositional information to the clusters increases the recall but the precision remains very low. The authors consider the approach useful for a preliminary step but only for a limited amount of words. The resulting structures obviously need human intervention for further refinement and structuring. It also becomes evident - because of the relatively low recall - that there are many UMLS concepts whose names are not formed by nominal strings or prepositional structures from the kind which is treated in [2, 7]. Please note that this well-documented work deals with ontology initiation using a minimum of preliminary available tools and resources; that is why we consider it very important and present it here in more details.

ASIUM, a more sophisticated system for concept acquisition, is presented in [8]. ASIUM learns knowledge from syntactically parsed text, so the input for the learning algorithm – conceptual clustering method - is fully analysed text. No concepts are given in advance. From the parsed input, ASIUM learns verb frames like:

```
<verb> <preposition | syntactic role: concept>.
```

A sample is given below:

```
<to drop> <object: Explosive> <in: Public_Place>
```

(the pairs `object:Explosive` and `in:Public_Place` are *subcategories*, `object` is a *syntactic role* and `in` is a *preposition* but `Explosive` and `Public_Place` are concepts used as *restrictions of selection*). The method learns by grouping similar subcategories into clusters. The resulting concepts are labeled by a human expert.

ASIUM relies on a language technology which is relatively sophisticated – a full syntactical analyser of French. The parser outputs also roles in the verb frames which look like the subcategorization frames but with concepts replaced by nouns:

```
<verb> <preposition | role: head noun>.
```

By grouping head nouns and semantic roles, ASIUM generalises the initial syntactic frames and covers by induction examples that did not occur as such in text. As shown in [8], ASIUM is a rather powerful concept acquisition system. Starting with the syntactic frames:

```
<to travel> <subject:[father,neighbour,friend]> <by:[car,train]>
<to drive> <subject:[friend,colleague]><object:[car,motorcycle]>
```

ASIUM will learn two concepts:

```
Human: father; neighbor; friend; colleague.
Motorized Vehicle: car; train; motorcycle.
```

and two subcategorization frames:

```
<to travel><subject: Human><by:Motorized Vehicle>
<to drive><subject: Human><object: Motorized Vehicle>
```

As the authors explain in [8], human experts control the link between the new concepts and the verbs because the threshold, fixed preliminary by the expert, does not avoid over-generalisation.

ASIUM inspired further research work for learning concepts from parsed texts. For instance [9] expands the ontology of the CIRCM-TUTOR system using a similar approach. It is difficult to compare directly all such systems for concepts acquisition, as they employ different modules for syntactic analysis and start from different inputs. But it is evident that (i) the deeper linguistic analysis has a positive effect to concept acquisition and (ii) the resulting concepts (clusters) always need to be revised by a human expert who also assigns them labels.

4 Learning Taxonomies and Enriching Ontologies

Here we briefly overview three approaches. One of them is based on deep linguistic information and very complex word sense processing. Its result is enriching

WordNet with new domain concepts. The second one combines patterns from different sources to build taxonomic relations. The third one processes the output of a parser and organises the words/concepts in a hierarchy, following the Formal Concept Analysis. In this section it becomes evident that the task of taxonomic structuring always exploits information from deeper text analysis, which provides additional evidence concerning the meaning of the words and concepts met in the texts.

The tool OntoLearn [10] extracts domain ontologies from documents shared among the members of virtual organizations. Its first step is to extract and filter the domain terminology from the available documents. Because of their low ambiguity and high specificity, these words or phrases are very good candidates to label domain concepts, as they denote important domain concepts and relations. There are well-known methods for extracting stable collocations from text. OntoLearn uses rule-based tools developed earlier by the team and extracts domain terms, which are further evaluated by two specific measures: *domain relevance* (how often the term appears in the domain corpus, compared to a larger collection of corpora) and *domain consensus* (which measures the distributed use of a term in a domain corpus). Combining both measures, OntoLearn computes the *weight* of each term. Accepted terms have weight over a threshold which is set experimentally. In this way, accepted terms for *tourism* are *travel information*, *shopping street*, *airline ticket*, *booking form*, etc. while for *finance* accepted terms are *vice president*, *net income*, *executive officer*, *composite trading* and so on.

After selecting the domain terminology, OntoLearn builds subtrees of terms (concept labels) according to simple string inclusion. For instance, an initial hierarchy of terms in the travel domain may look like the one shown at Fig. 2. Without semantic

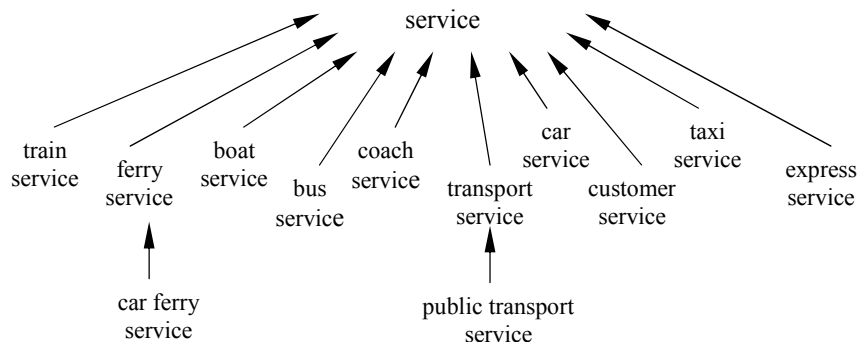


Figure 2. A lexicalized tree in the domain of tourism [10]

interpretation, Fig.2 contains erroneous classifications, e.g. *bus service* is not classified as *public transport service*. The main innovation of OntoLearn is that it performs semantic interpretation of the complex terms by constructing it compositionally. OntoLearn uses the general WordNet senses and appropriate conceptual relations that hold among the concept components. The semantic resources are processed by the so-called *structural semantic interconnection* algorithm, a novel knowledge-based iterative approach to word sense disambiguation. Very roughly, as

an illustration only, we remind that in WordNet *bus* is a kind of (*public*) *transport*, so it is possible to compute that *bus service* has to be classified as a *public transport service*. In this way the complex domain terms can be semantically interpreted and arranged in a hierarchical manner (please note that the terms – single words have to be defined somehow, either in WordNet or in thesauri, otherwise OntoLearn has no way to calculate their meaning). Figure 3 shows a domain concept tree, obtained from the lexicalized tree after the semantic interpretation. We see there that an *express* can be a *bus* or a *train*, and both interpretations are valid because they are obtained from relations between terms within the domain. Conceptual relations play important role in the semantic interpretation. The chosen kernel of conceptual relations is built using the definition of basic relations given in [11].

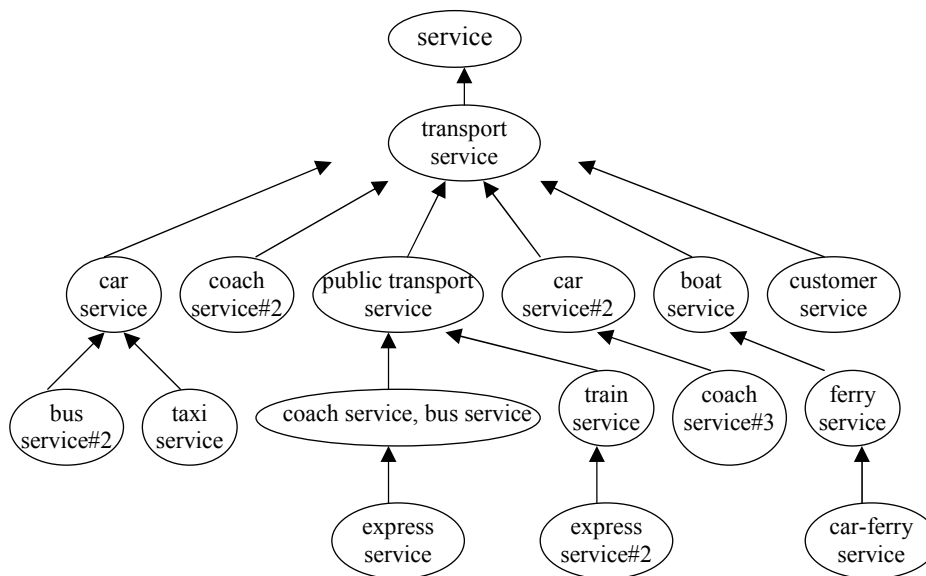


Figure 3. Domain concept tree [10]

OntoLearn evaluation is difficult as the ontology assessment is still an open problem. There may be experimental validation and evaluation, by users who apply the ontology, as well as assessment against methodological evaluation guidelines. OntoLearn was evaluated as a tool supporting ontology engineers in the Harmonise project. During the first year of the project, a core ontology with 300 concepts was manually defined. Simultaneously OntoLearn extracted an initial list of 14,383 candidate terms and organised them in a domain concept forest of 3,840 concepts. The latter resource was submitted to the domain experts for ontology updating and integration. The obtained precision ranged from 72.9% to about 80% and the recall was 52.74%. The precision shift is due to the well-known fact that experts may have different intuitions about the relevance of a concept. The authors conclude: "...In any case, OntoLearn favoured a considerable speed up in ontology development, since shortly the Harmonise ontology reached about 3,000 concepts. Clearly, the definition

of an initial set of basic domain concepts is crucial, so as to justify long lasting and even harsh discussions. But once an agreement is reached, filling the lower levels of the ontology can still take a long time simply because it is a tedious and time-consuming task.” Therefore the authors consider OntoLearn as a very useful tool within the Harmonise project [10]. A drawback of OntoLearn is that it is unable to analyze totally unknown terms as it exploits much linguistic information which is available for the words known to the system only.

In the computational linguistics, OntoLearn is a state-of-the-art result for ontology construction using semantic interpretation. The aim is to enrich a general purpose ontology – WordNet – with the detected domain concepts (as WordNet does not cover complex domain terms). Today only few papers propose to enrich existing ontologies with new concepts. Another approach addressing documents in Internet is reported in [12], where the authors collect for each concept in WordNet the words that appear most distinctively in texts related to it (*topic signatures*) and apply them for clustering the concepts that lexicalize the word senses of a given word.

The results reported in [13] illustrate the spirit of the information retrieval approaches to taxonomy learning (in contrast to the rule-based solution of OntoLearn presented above). The system described in [13] builds taxonomic links by solving a classification tasks for every concept pair. Let us consider two terms, say *conference* and *event*. They could be either unrelated, or taxonomically related in three different ways: *is-a(conference,event)*, *is-a(event,conference)*, *siblings(conference,event)*. Therefore, intuitively, the decision is to gather as many different sources of evidence as possible and choose the relation with maximal evidence according to all of them. This approach combines information from:

- (i) Linguistic patterns explicating *is-a* relations matched to a large text corpus [14];
- (ii) Linguistic patterns explicating *is-a* relations matched to the Web [14];
- (iii) WordNet and its hyponyms, and
- (iv) the lexicalized *is-a* relation, as exemplified at Fig. 2.

For instance, a pattern for searching hyponyms for NounPhrases (NP) in corpus is NP0 such as NP1, NP2, ..., NPn-1 and/or NPn (see [14] for other patterns).

Similar patterns for searching in Internet may reveal descriptions of hyponymic relations. Table 1 summarises some results for *is-a* and the related probabilities:

<i>conference is-a</i>	<i>conference is-a</i>
<i>event</i> 0.44	<i>service</i> 0.27,
<i>meeting</i> 0.11	<i>meeting</i> 0.11,
<i>activity</i> 0.11	<i>activity</i> 0.11
A. Extraction from corpus	B. Extraction from Internet

Table 1. Evidence for taxonomic relations, extracted from different sources

The four sources of evidence are normalized in order to be comparable and combined by two very simple techniques (to take the *mean* of all possible values and to use the *maximum*). The result shows that combining diverse and heterogeneous

information indeed leads to better results than classification according to a single source (and *conference is-a event*). The approach is evaluated against a handcrafted ontology for the touristic domain. The *mean* strategy with all WordNet senses yielded precision 17.16% and recall 29.84% with threshold $t=0.01$; and the one with the first WordNet senses – precision 17.38% and recall $R=29.24\%$ at $t=0.01$. The *max* strategy with all senses yielded a precision 16.03% and recall 29.87% at $t=0.04$. These results are comparable and show that there is a lot of potential in the combination of different approaches for text extraction.

The last result which we overview is presented in the paper [15]. It builds a taxonomy starting from the results of a parser. We explain the construction by example. Let us consider the sentences:

People book hotels. The man drove the bike along the beach.

After parsing them, and after identifying the basic word forms (which turns *hotels* to *hotel* and *drove* to *drive*), the following semantic relations can be extracted from the parsing results:

book_subj(people) drive_obj(bike) drive_subj(man)
 book_obj(hotel) drive_along(beach)

Not all dependencies from this kind are interesting, but the repeating ones reveal semantic links between the verb, the thematic role and the filler (word or concept). Thus *hotels* are *bookable*, *bikes* are *drivable*, *men* are *driving* etc. And in contrast to the similarity-based clustering, which is popular in taxonomy construction, one can consider the Formal Concept Analysis [16] as an alternative set-theoretic classification. The authors show a taxonomy built for the tourism domain, using the attributes as presented in the example above (Table 2 and Figure 4). The claim is that Figure 4 is much easier to read and trace for the human engineer, as in keeps the labels of the classification features. The conceptual hierarchy is built on the basis of the inclusion relations between the selectional restrictions of all the verbs. Thus FCA supports tracing of reasons why a taxonomy looks the way it is, according to linguistic knowledge acquired from text. Figure 4 is in fact an interesting idea of linking natural language processing and FCA in order to trace the properties of the extracted knowledge chunks. If the input corpora are updated regularly, there is a natural way to justify the ontology evolution accordingly. A more detailed evaluation of this approach is presented in [17].

	bookable	rentable	driveable	rideable	joinable
apartment	X	X			
car	X	X	X		
motor-bike	X	X	X	X	
excursion	X				X
trip	X				X

Table 2. Concepts and attributes as a formal context in the domain of tourism

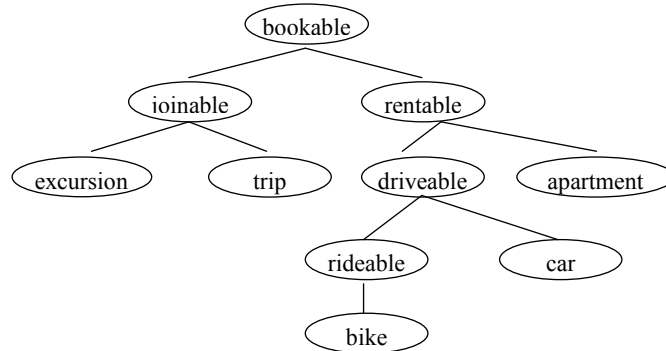


Figure 4. A taxonomy built for Table 2, which reflects features extracted by parsing.

5 Discussion and Conclusion

In this paper we summarise typical approaches of how to employ language technologies in order to acquire knowledge from texts. We prefer to focus on prototypes which clearly organize the extracted language units into conceptual structures. For clarity and brevity, we grouped the approaches into three main categories: acquisition of instances, concepts, hierarchies. We emphasize that the extracted structures are often domain-dependent (because they are based on the input corpus) and need further refinement and formalization.

Despite the variety of the particular applications, there are some established strategies of how to apply language technologies in knowledge acquisition. For instance, it is common to extract phrasal descriptions after parsing and to encode the corresponding knowledge chunks. The assumption is that verbs pose more or less strong selectional restrictions on their arguments and this linguistic phenomenon reflects semantic relationships. The better the parser is, the more sentence phrases it extracts and the final result is more sophisticated and more accurate. However, the relatively low recall of this task is an evidence that many concepts are not communicated via compact phrasal descriptions and therefore are difficult to grasp by automatic text processing.

Together with the main language technologies in use (lexical and morphological analysers, POS taggers, tools for recognition of named entities, parsers, extractors of collocations, semantic interpreters of word senses) we exemplified the main classification technology applied for building a taxonomy – the unsupervised clustering. Mostly the *is-a* relation is considered. The resulting ontological structures are not perfect but they are viewed as drafts that need further justification by human experts. The concepts in the taxonomies mirror the words (strings) occurring in the source texts and the human expert may choose to delete some of them. If a word has several meanings and the input corpora are really large, then mining by patterns or counting the verb-object collocations will hopefully reveal all senses as shown at

Table 1. It is still unclear how to evaluate the usefulness and the correctness of a particular ontology acquired for application in certain domain.

However, word senses are *not* ontological concepts although there are many similarities between lexicons and ontologies [18]. An ontology is a set of categories and relationships among them while a lexicon depends on the word senses in particular natural language. In the same way the text semantics, extracted in formal propositions, is not conceptual model of the information encoded in the text. Moreover, gathering facts from Internet, we have no methods to evaluate whether the extracted knowledge is relevant to the real world. For instance, the NER module might collect named entities from science fiction sites. So even the results of the successful NER task, which are extracted with more than 90% linguistic accuracy, need conceptual verification before asserting them into a formal conceptual model. From this perspective, extracting knowledge from dedicated domain sites (e.g. ePortfolio portals) is a more promising approach for automatic ontology construction.

We have already said that attempts for automatic extraction of logical propositions are rare as the complexity of the task is well-known. Extraction of one fact from two sentences is not trivial, especially using the methods applied today, as the current parsers work over single sentences only. Another ontological construct that is not tackled automatically is the description of the classification perspective, since it is rarely encoded as a compact statement in free texts.

One may argue that the results presented above are discouraging because of their low accuracy. However the author prefers the optimistic vision. All the prototypes are at their initial implementation stage and the cited papers report work in progress. Moreover, the quality of the language technology tools is growing incrementally and this contributes to the increasing quality of the very large linguistic resources. In addition, there are many publicly available tools (e.g. POS-taggers and parsers for English) and open resources like WordNet. The interest in the multilingual resources and tools is growing too and much bilingual terminology is already extracted automatically. This facilitates the work of research groups in smaller countries who deal with languages other than English. The author believes that soon there will be large open archives of electronic text resources in a number of languages and many publicly available tools for language processing (similarly to the paper archives which were collected also incrementally). All these stimuli motivate further research attempts in automatic ontology acquisition. It seems unlikely that the quality of the automatically extracted ontologies will be satisfactory enough but at least drafting ontologies will be much easier. The author also believes that methods for ontology assessment against domain corpora will be developed in the near future.

REFERENCES

- [1] T. Berners-Lee. *Weaving the Web*. Harper, 1999.
- [2] Reinberger, M.-L., Spyns, P., Pretorius, A.J. and Daelemans, W. *Automatic Initiation of an Ontology*. In R. Meersman, Z. Tari (Editors), Proc. CoopIS/DOA/OBDASE 2004, Springer, Lecture Notes in Computer Science 3290, 2004, pp. 600-617.

- [3] L. Hirschman. *The Evolution of Evaluation: Lessons from the Message Understanding Conferences*. Computer Speech and Language 12, 1998, pp. 281-305.
- [4] Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D. and Kirilov, A. *KIM – a Semantic Platform for Information Extraction and Retrieval*. Journal of Natural Language Engineering 10 (3/4), 2004, pp. 375-392, see also URL: <http://62.213.161.156/KIM/screen/KWUIMain.jsp>
- [5] Handschuh, S. and S. Staab. *CREAM: CREATing Metadata for the Semantic Web*. Computer Networks: The International Journal of Computer and Telecommunications Networking, Volume 42, Issue 5, 2003, pp. 579 – 598.
- [6] Maedche, A. and S. Staab. *Ontology Learning for the Semantic Web*. IEEE Intelligent Systems 16 (2), Special Issue on Semantic Web, 2001, pp. 72-79.
- [7] Reinberger, M.-L. and P. Spyns: *Discovering Knowledge in Texts for the Learning of DOGMA-inspired Ontologies*. In the Proc. ECAI-2004 Workshop on Ontology Learning and Population: *Towards Evaluation of Text-based Methods in the Semantic Web and Knowledge Discovery Life Cycle*, August 2004, pp. 19-24.
- [8] Faure, D. and T. Poibeau. *First Experiments of Using Semantic Knowledge Learned by ASIUM for Information Extraction Task Using INTEX*. In the Proc. of the Workshop on Ontology Learning, ECAI 2000, pp. 7-12.
- [9] Lee, C.H., Seu, J. H. and M. Evens. *Building an Ontology for CIRCSIM-Tutor*. Proc. 13th Midwest AI and Cognitive Science Society Conference, MAICS-2002, Chicago, pp. 161-168.
- [10] Navigli, R. and P. Velardi. *Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites*. Journal of Computational Linguistics, Vol. 30, Issue 2, June 2004, pp. 151 - 179.
- [11] Sowa, J. *Conceptual Structures: Information Processing in Mind and Machine*, 1984, Addison-Wesley, Reading, MA.
- [12] Agirre E., Ansa1, O., Hovy E. and D. Martínez. *Enriching very large ontologies using the WWW*. In the Proc. of the Workshop on Ontology Learning, ECAI 2000, pp. 37-42.
- [13] Cimiano P., Pivk A., Schmidt-Thieme L. and S. Staab. *Learning Taxonomic Relations from Heterogeneous Evidence*. In the Proc. ECAI-2004 Workshop on Ontology Learning and Population, 2004.
- [14] Hearst, M.A., *Automatic Acquisition of Hyponyms from Large Text Corpora*, in Proc. COLING-1992, pp. 539-545.
- [15] Cimiano P., Hotho, A. and S. Staab. *Comparing Conceptual, Divisive and Agglomerative Clustering for Learning Taxonomies from Text*. In the Proc. ECAI 04, IOS Press, 2004, pp. 435-439.
- [16] Ganter, B. and R. Wille, *Formal Concept Analysis – Mathematical Foundations*, Springer Verlag, 1999.
- [17] Cimiano, P., Staab, S. and J. Tane. *Automatic Acquisition of Taxonomies from Text: FCA meets NLP*. In the Proc. of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining, Cavtat--Dubrovnik, Croatia, 2003, pp. 10-17.
- [18] Hirst, G. *Ontology and the lexicon*. In: Staab, S. and R. Studer (Eds.), *Handbook on Ontologies*, Berlin: Springer, 2004, pp. 209-229.