Ontology Building Using Parallel Enumerative Structures

Mouna Kamel
Institut de Recherche en Informatique de Toulouse
(IRIT) – CNRS – UPS,
118, Route de Narbonne, 31062 Toulouse, France
(+33) 5 61 55 83 38
kamel@irit.fr

Bernard Rothenburger
Institut de Recherche en Informatique de Toulouse
(IRIT) – CNRS – UPS
118, Route de Narbonne 31062 Toulouse, France
(+33) 5 61 55 83 38
rothenburger@irit.fr

ABSTRACT

The semantics of a text is carried by both the natural language it contains and its layout. As ontology building processes have so far taken only plain text into consideration, our aim is to elicit its textual structure. We focus here on parallel enumerative structures because they bear implicit or explicit hierarchical relations, they have salient visual properties, and they are frequently found in corpora. We have defined a process which identifies them in a text, translates them into ontology structures and finally links such structures to the concepts of an existing ontology. We have assessed this process on Wikipedia encyclopaedic articles as they are rich in definitions and statements, and contain many enumerations. The many ontology structures we have obtained are thus used to enrich an ontology which we had automatically built from database specification documents.

Categories and Subject Descriptors

I.2.7 Natural Language Processing - *Text analysis,* I.2.6 Learning - *Knowledge acquisition*

General Terms

Algorithms, Documentation, Languages

Keywords

Ontology building and enrichment from text, layout analysis, NLP tools.

1. MOTIVATION

Many approaches have been suggested for the construction, enrichment or population of ontology from text. They are based on lexical, syntactical, semantic or rhetorical aspects of natural language. They encompass machine learning [1], specific natural language processing tools [2], or combination of both [3]. These methods are usually applied on plain texts. However, a large variety of layouts or structures can be found in the visual presentation of a text with a diversity of interpretations for each of them [4]. Some of them implicitly carry ontological knowledge as shown in example 1. The meaning carried by this structure may be expressed through the sentence in example 2. In both cases, a human being may easily deduce the conceptual framework presented in figure 1.

In the case of sentence analysis (example 2), the automatic deduction by a Natural Language Processing (NLP) tool of its formal counterpart is a very tricky issue which will necessitate to carry out non trivial tasks such as the resolution of anaphora or the design of sophisticated multi-sentence textual patterns.

> *Under IAU definitions, in the Solar System and in order of increasing distance from the Sun, there are eight planets:*
>
> - *four terrestrials:*
>   - *- Mercury,*
>   - *- Venus,*
>   - *- Earth,*
>   - *- Mars.*
> - *four gas giants:*
>   - *- Jupiter,*
>   - *- Saturn,*
>   - *- Uranus,*
>   - *- Neptune.*

Example 1 : a structure which carries ontological knowledge

> *Under IAU definitions, there are eight planets in the Solar System. In order of increasing distance from the Sun, they are the four terrestrials, Mercury, Venus, Earth, and Mars, then the four gas giants, Jupiter, Saturn, Uranus, and Neptune.*

Example 2 : a sentential representation of the example 1

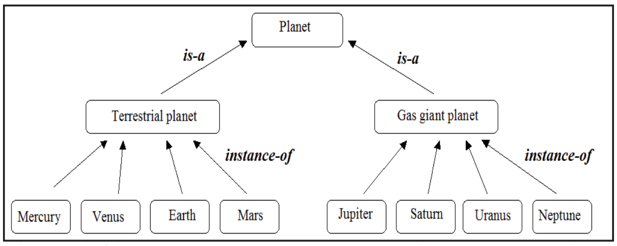

Figure 1. Conceptual network corresponding to the meaning of examples 1 and 2

However for layout structure analysis (example 1), different parts of the knowledge are more easily identifiable thanks to lexical or typo-dispositional marks. We claim that it becomes thus easier to identify in an automated way the corresponding conceptual network. The above meaning-bearing layouts allow a straightforward identification of ontological relations: often hyperonymy, sometimes meronymy, and occasionally other relations.

We focus here on a specific kind of meaning-bearing layout that we call parallel enumerative structures (PES). Example 1 is typical of such a layout. These structures present some regularities and appear very frequently. Their analysis could be a relevant contribution to improve knowledge elicitation and modelling from text. Moreover, it would provide new triggers for the identification of new concepts or semantic relations, therefore enabling to go beyond the classical ontology learning approaches which only consider the plain text.

2. TRANSLATION PROCESS

An *enumeration* is a set of items with or without semantic relations between them. An *item* is a co-enumerated entity which can be discernable by typographic, dispositional and/or lexico-syntactic marks. And a *parallel enumeration* is a paradigmatic enumeration (*i.e.* all items are functionally equivalent, textually or syntactically), visually homogeneous (*i.e.* all items are visually equivalent) and isolated (*i.e.* no item is linked to any textual unit which is out of the enumeration). An *introductory phrase*, hereafter called *primer,* is a phrase or a sentence which introduces an enumeration, and which is identifiable by lexico-syntactic and/or typo-dispositional marks. Finally, let us call *parallel enumerative structure* (PES) a vertical textual structure composed of a primer and a parallel enumeration.

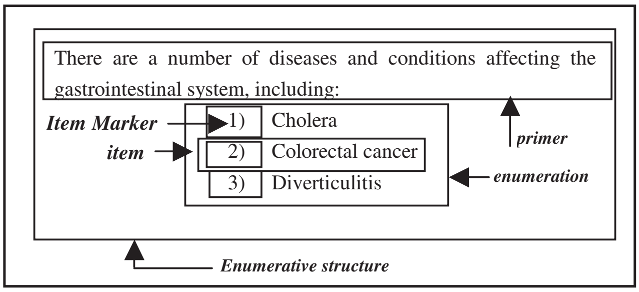

Figure 2. Composition of an enumerative structure

Broadly speaking, the idea is to translate a PES into a single ontology structure (i.e. one or two-level hierarchy) according to the following principles: (1) the primer contains one father concept and one semantic relation which links this father concept to concepts contained in the items, (2) each item contains one child concept semantically related to the father concept of the primer, (3) all child concepts will be considered as belonging to the same conceptual level. An example of this correspondence is the structure obtained in Figure 1 from the example 1.

The syntactic structure of the primer helps to identify the father concept and the semantic relation it contains. We have characterized 3 cases:

→ *The primer is not syntactically correct.*

- The primer could be composed of a noun phrase. This noun phrase represents the father concept and the semantic relation is the relation *is-a*.

- The primer ends with a verb phrase at the active form. The semantic class to which this verb belongs reflects the nature of the relation and the father concept corresponds to the main term of the noun phrase which is the subject of this verb.

→ *The primer is complete*. It contains a lexical unit taken from a gazetteer or a number which specifies the number of items. The concept father is the term which co-occurs with this lexical marker, and the relation is the relation *is-a*.

→ *The primer is syntactically correct and not complete*. The father concept may be found in the subject noun phrase or in the object noun phrase of the main clause and may be eventually detected thanks to heuristics. The relation is the relation *is-a*.

Our method consists in (1) identifying each enumerative structure and its different components (primer and items), (2) checking whether the enumeration is parallel, (3) identifying the father concept and the nature of the semantic relation, (4) extracting the child concepts from each item and (5) building an ontological structure. This fifth step is based on annotations produced over the four previous steps.

3. APPLICATION

Wikipedia documents are encyclopaedic and contain a lot of definitional statements and properties. Furthermore, articles are written according to a comprehensive set of editorial and structural guidelines. Actually it thus advocates the writing of PES. The experiment reported in this paper concerns the enrichment of an existing ontology which is a frame of reference used to localise information relating to urbanism, environment and territorial organisations. It contains both geographical and real-world concepts. This ontology has 728 concepts. We then obtain 182 disambiguated pages which contain at least one PES (according our criteria). From these 182 articles we exploit 276 PES which allowed to enrich our ontology with 349 new concepts and 201 instances which were considered as relevant by experts and knowledge engineers involved in the building of this ontology.

4. FUTURE WORKS

In the short-term, our idea is to combine our approach with the usual ontology learning from text ones. For example, in order to better take advantage of Wikipedia's articles, it would seem interesting to complete the approach of Herbelot et al. [5], which exploits plain text only. We also plan to exploit redirect links and homonym pages to maximise the number of relevant articles. On the other hand we want to improve the analysis of enumerative structures by going beyond simple parsing, particularly regarding the primer. Authors may use complex grammatical constructions or linguistic variations in their writing, even within the enumerative structures. We then face problems of anaphora resolution, ellipses, apposition, extraposition and rhetorical forms, etc. Also, discourse analysis must be carried out to process non-parallel enumerative structures.

5. REFERENCES

[1] Nédellec, C., Nazarenko, A.: Ontology and Information Extraction. *in S. Staab & R. Studer (eds.) Handbook on Ontologies in Information Systems*, Springer (2003)

[2] Giuliano, C., Lavelli, A., Romano, L.: Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In Proc. EACL (2006)

[3] Giovannetti, E., Marchi, S., Montemagni, S.: Combining Statistical Techniques and Lexico-syntactic Patterns for Semantic Relation Extraction from Text. Fifth workshop on Semantic Web Applications and Perspectives, FA0-UN, Roma, Italy (2008).

[4] Virbel, J., Luc, C.: Le modèle d'architecture textuelle: fondements et expérimentation. *Verbum*, Vol. XXIII, N. 1, p. 103-123 (2001)

[5] Herbelot, A., Copestake, A., 2006: Acquiring ontological relationships from Wikipedia using RMRS. In: Proceedings of the International Semantic Web Conference 2006. Workshop on Web Content Mining with Human Language Technologies, Athens, GA (2006).