ONTOLOGY BUILDING USING PARALLEL ENUMERATIVE STRUCTURES

Mouna Kamel and Bernard Rothenburger
Institut de Recherche en Informatique de Toulouse (IRIT) – CNRS
UPS, 118, Route de Narbonne, 31062 Toulouse Cedex, France

Keywords: Ontology building and enrichment from text, Layout analysis, NLP tools.

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 ontological 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 ontological structures we have obtained are thus used to enrich an ontology which we had automatically built from database specification documents.

1 INTRODUCTION

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 tools (Nédellec and Nazarenko, 2003), specific natural language processing tools (Giuliano et al., 2006), or combination of both (Giovannetti et al., 2008). 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.

Example 1: a structure which carries ontological knowledge.

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.*

Some of these structures implicitly carry ontological knowledge as shown in the example 1. The meaning carried by this structure may be expressed through the example 2:

Example 2: a sentential representation of the example 1.

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.

In both cases, a human being may easily deduce the following conceptual framework:

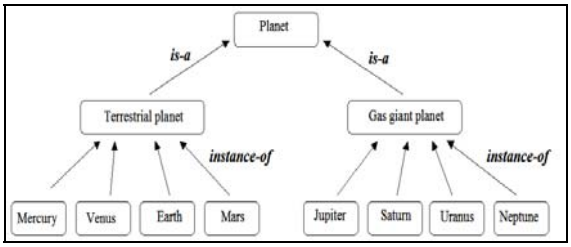

Figure 1: Conceptual network corresponding to the meaning of examples 1 and 2.

In the case of sentence analysis (example 2), the automatic deduction by a Natural Language Processing (NLP) tool of its formal counterpart is a

Kamel M. and Rothenburger B..
ONTOLOGY BUILDING USING PARALLEL ENUMERATIVE STRUCTURES .
DOI: 10.5220/0003097602760281
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2010), pages 276-281
ISBN: 978-989-8425-29-4

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. 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 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.

This paper is organized as follows. Section 2 recalls the role and the importance that the layout and the structure may have in the textual semantics. Section 3 gives an analysis of parallel enumerative structures and details the process we define to translate a parallel enumerative structure into an ontological structure. And in section 4 we describe an application whose aim is to enrich an existing ontology through applying our translating process on Wikipedia articles which contain many parallel enumerative structures, thus evaluating our translation process. Finally we draw the balance of our current work and list several future advancements required to go further.

2 STRUCTURE AS PART OF TEXTUAL SEMANTICS

When producing a document, a writer may use not just linguistic skills but also the ability to logically and physically structure his/her writing. In this regard the materiality of a text is part of its meaning: "the overlay of graphics on text is in many ways equivalent to the overlay of prosody on speech" (Power et al., 2003). So the document structure, also termed page layout, has been investigated for different purposes, essentially for the improvement of text generation (Mann et al., 1992), (Power et al., 2003) and text segmentation systems (Hearst, 1997).

On the other hand, as documents are increasingly available in a digital format, their contents become easily accessible thanks to the existing mark-up languages as XML whose tags convey the semantics of the structure. These tags have been particularly taken into consideration to improve text summarization systems, as in (Groza et al., 2007) or to perform Web pages classification (Shen et al., 2007). (Auer et al., 2007) transform preformatted tables into a set of triples (subject, predicate, object) which are then stored in a database. We have also already formulated a procedure based on the semantics of tags and on their nested levels to build an ontology kernel from a collection of structured documents (Kamel and Aussenac-Gilles, 2009).

Textual objects are “textual segments that correspond to a specific metalinguistic formulation which is highlighted by a specific layout” (Rebeyrolle and Péry-Woodley, 1998). One textual object which aroused great interest is the enumeration because it visually emphasizes information and encompasses a concept or an idea which is specified into various elements for a summarization purpose. Moreover it maintains relationships between its different components, presents regularities and occurs quite frequently. Discursive enumerations (also termed *in-line list* or *horizontal enumeration*) are distinguished from vertical enumerations by the way they are written. They are indicated by lexico-syntactic marks which may induce an ambiguous meaning within the sentence (example 2). Vertical enumerations are indicated by salient visual and typo-dispositional marks which facilitate reading comprehension. Although the elements of these enumerations are visually discontinuous, they constitute a whole at the semantic level (example 1). Their identification and their interpretation in a text are regular enough to be automatically computed. Actually, Luc led a study on enumerations, proving that there is a functionally equivalence between discursive enumerations and vertical ones (Luc, 2001).

To our best knowledge, only a few works have analysed the layout of a document to build linguistic resources (Jacquemin and Bush, 2000), but none for the ontology building process. We propose here to show the benefits we have obtained by the analysis of the enumerative structures for such a process.

3 THE 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. Then we can distinguish:

- A *parallel enumeration* as 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) (fig. 3-a),
- A *Non-parallel enumeration* as an enumeration missing one or more properties of a parallel enumeration (fig. 3-e).

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.

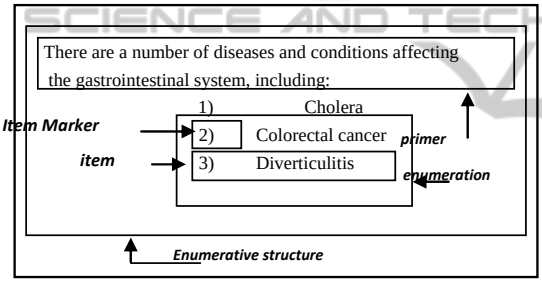

Figure 2: Composition of an enumerative structure.

Broadly speaking, the idea is to translate a PES into a single ontological 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 p 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 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 (fig. 3-b). This noun phrase represents the father concept and the implicit semantic relation is the relation *is-a*.

- The primer could end with a verb phrase at the active form (fig. 3-a). 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* (fig. 3-c). 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 implicit relation is the relation *is-a*.

→ *The primer is syntactically correct and not complete* (fig. 3-d). 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 implicit 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.

4 EVALUATION

We carry out an experiment which will estimate the enrichment ratio of an existing ontology when exploiting PES from Wikipedia pages.

4.1 Experiment Setup

Within the GEONTO (http://geonto.lri.fr/) project, ontologies are automatically built from structured database specifications documents (given in an XML format). To enrich these ontologies, we use Wikipedia documents. Wikipedia documents are encyclopaedic: each article describes a single concept, and there is a single article for each concept. This is an interesting feature because it will avoid having to cope with the problem of polysemy. These articles contain a lot of definitional statements

Fig 3-a: article *Language* cf. Wikipedia.	**Fig 3-b**: article *Antihypertensive drug* cf. Wikipedia.
Non-spoken forms of communication are : ▪ Written language ▪ Sign language ▪ Whistled language ▪ Drum language ▪ Non-verbal language	Aldosterone receptor antagonists: • Eplerenone • Spironolactone
Fig 3-c: article *Library classification* cf. Wikipedia.	**Fig 3-d**: article *Flora* cf. Wikipedia.
As a result of differences in Notation, history, use of enumeration, hierarchy, facets, classification systems can differ in the following ways : ▪ Type of Notation: Notation can be pure (consisting of only numerals for example) or mixed (consisting of letters, numerals, and other symbols). Expressiveness: This is the degree in which the notation can express relationship between concepts or structure.	Lastly, floras may be subdivided by special environments: ▪ Native flora. The native and indigenous flora of an area. ▪ Agricultural and garden flora. The plants that are deliberately grown by humans. ▪ Weed flora. Traditionally this classification was applied to plants regarded as undesirable ….
Fig 3-e: article *Library classification* cf. Wikipedia.	
The justifications for this protocol are: ▪ Children, especially the younger ones, have normally not yet developed the mental capacity to fully comprehend the hazards ▪ It is impractical in many cases to avoid children crossing the traveled portions of roadways after leaving a school bus or to have an adult accompany them. ▪ The size of a school bus generally limits visibility for both the children and motorists during loading and unloading.	

Figure 3: Examples of enumerative structures.

and properties. Furthermore, articles are written according to a comprehensive set of editorial and structural guidelines. For bulleted and numbered lists, the Manual of Style (http://en.wikipedia.org/ wiki/ Wikipedia:Manual_of_Style). recommends : "use the same grammatical form for all elements in a list, and do not mix the use of sentences and sentence fragments as elements". Actually it thus advocates the writing of PES. So such texts are a goldmine for the mining of meaning (Medelyan et al., 2009). Several others works exploit Wikipedia documents for information extraction process or ontology building process. Some of these works are based on a pattern matching approach improved by more efficient parsers (Herbelot and Copestake, 2006), selectional restrictions (Wang et al., 2007) or anaphora resolution (Nguyen et al., 2007). Other works have essentially focused on the exploitation of categories to extend taxonomies (Chernov et al., 2006), and on infoboxes to populate the RDF triples DBpedia dataset (Auer et al., 2007).
Our approach is different since it takes advantage of the textual structure of these Wikepedia documents to improve the ontology building process.

In order to perform the translation process described in section 4, we have implemented a text processing chain using the GATE NLP platform (http://gate.ac.uk). This platform allows the development of pipeline processes which run a set of NLP tools and may use linguistic resources. Each step adds new annotations to the corpus, even sometimes using annotations previously set. An annotation steps may also use Java Annotation Patterns Engine (JAPE) rules. JAPE is a language which allows to define regular expressions over annotations.

The experiment reported in this paper concerns the enrichment of the OntoBDTopo ontology which was built from the BDTopo database specifications (one of the database used by GEONTO project). BDTopo is a frame of reference used to localise information relating to urbanism, environment and territorial organisations. It contains both geographical and real-world concepts. The OntoBDTopo ontology has 728 concepts.

4.2 Experiment Results

We first leave aside:

- Parallel enumerative structures whose primer does not end with a colon (these have specificities which made them out of the scope of this study),

Table 1: Experiment results.

Feature	Number	Comment
Initial concepts (IC)	728	
Total Wikepedia pages (TWP)	406	55% of IC lead to a Wikipedia entry
Usable Wikipedia pages (UWP)	283	39% of IC lead to a disambiguated Wikipedia page
Usable pages with PES (UWPES)	182	25% of IC contain at least one PES
PES (PES)	434	The total number of PES present in the 182 UWPES
Usable PES (UPES)	276	64% of PES are relevant for the translation process
Complete PES (CPES)	52	Concern 19% of PES
Correct and non-complete PES (CNCPES)	149	Concern 53% of PES
Non-correct PES (NCPES)	71	Concern 26% of PES
New concepts (NC)	349	48% of IC
New instances (NI)	201	Populate the ontology

- Standard appendices such as "*see also*", "*other wikimedia projects*", etc. which do not bear ontological structure.

We then obtain 182 disambiguated pages which contain at least one PES (according to 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 OntoBDTopo specification . Table 1 details these results.

4.3 Results Discussion

Ontology quality assessment is a multifaceted problem. It can be based on quantitative measures (proximity to another ontology, coverage by a corpus, quality of search results, etc.) or on qualitative aspects (logical consistency, conceptual validity, expert validation, etc.). But there is no “gold standard” evaluated by experts that could be used as a reference against the ontology we produce automatically. Here, we have chosen to estimate the number of new relevant concepts and relations our translation process adds to an existing ontology. We observe than the number of concepts goes up by 50%. Concerning relations and according to the above typology of primers, we have identified more than 80% of taxonomic relation and 15% of meronymic ones. The few remaining ones are other relations (mainly issued from NCPES) which we can identify in the primer by NLP tools.

Since the root concept label of an ontological structure we get is already a label (or includes a label) of the original ontology, we carry a fusion of this latter with the new structure. This approach has the advantage of being fully automatic. Nevertheless, we are aware that the fusion process in turn carries specific problems which are out of the scope of the present article.

5 CONCLUSIONS

The aim of this study is to show that the structure of a text may play an important role in the ontology building process. Textual objects such as titles, enumerative structures, definitions, etc. which own important visual properties, bear implicit or explicit semantic relations between the concepts they contain. We have chosen to analyse parallel enumerative structures because of their salient visual properties and because they convey ontological properties. In fact, they express that a same semantic relation, expressed in the primer, links one concept in the primer to one concept in each item. Primer and items may be identified by typo-dispositional marks. On the other hand, most writing tools facilitate the layout, and make it that we increasingly find, amongst others, enumerative structures in electronic documents.

After noting that the understanding and the interpretation of an enumerative structure depends on the type of the primer (its lexical and syntactic properties), we have defined strategies

based on the analysis of this primer to translate a parallel enumerative structure into an ontological structure. The translation process consists in exploiting successive annotations set in the text with the help of the above mentioned NLP tools and rules.

We decided to assess our method within an ontology enrichment context. We show that the sole extraction of the information carried by the parallel enumerative structures improves significantly an existing ontology in terms of domain coverage.

In the short-term, the idea is to combine our approach to the usual ontology learning from text ones. In this way, so as to better take advantage of Wikipedia's articles, it would seem interesting to complete the approach of (Herbelot and Copestake, 2006) which exploits plain text only. We plan to also exploit in this context 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. (fig. 1.). Also, discourse analysis must be carried out to process non-parallel enumerative structures.

REFERENCES

Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiac, R., Ives, Z., 2007. DBpedia : a nucleus for a web of open data. *In: Proceedings of the Sixth International Semantic Web Conference and Second Asian Semantic Web Conference (ISWC/ASWC2007)*, Busan, South Korea, vol. 4825, pp 715-728

Chernov, S., Iofciu, T., Nejdl, W., Zhou, X. , 2006. Extracting semantic relationships between Wikipedia categories. *In: Proceedings of the First International Workshop : SemWiki'06* - From Wiki to Semantics. Co-located with the Third Annual European Semantic Web Conference ESWC'06 in Budva, Montenegro

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)

Groza, T., Handschuh, S., Möller K., Decker, S., 2007. SALT - Semantically Annotated LaTeX for scientific publications. *In: Proceedings of the 4th European Semantic Web Conference (ESWC 2007)*. Innsbruck, Austria

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

Hearst M. A.: TextTiling, 1997. Segmenting Text into Multi-paragraph Subtopic Passages. Computational Linguistics, volume 23, Number 1

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

Jacquemin C., Bush C., 2000. Fouille du Web pour la collecte d'Entités Nommées. In : E. Wehrli (Ed.), TALN 2000, Lausanne

Kamel, M., Aussenac-Gilles, N., 2009. *How can document structure improve ontology learning?* (regular paper). In: *Semantic Authoring, Annotation and Knowledge Markup Workshop - collocated with K-CAP 2009 (SAAKM 2009)*, Redondo Beach, California (USA), Siegfried Handschuh, Michael Sintek (Eds.), CEUR Workshop Proceedings, p. 1-8

Luc, C., 2001. Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte. TALN2001, Université de Tours, p. 263-272

Mann, W. C., Matthiessen, C. M., Thompson, S. A., 1992. Rhetorical structure theory and text analysis. In: Mann, W. C. and Thompson, S. A., editors, Discourse Description, Diverse Linguistic Analyses of a Fund-Raising Text, pp. 39-78. John Benjamins publishing Compagny, Amsterdam/Philadelphia

Medelyan O., Milne D., Legg C., Witten I.H., 2009. Mining meaning from Wikipedia. International *Journal of Human-Computer studies*. Volume 67, Issue 9, pp.716-754

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

Nguyen, D.P.T., Matsuo, Y., Ishizuka, M., 2007. Relation extraction from Wikipedia using subtree mining. In: *Proceedings of the AAAI'07 Conference*, Vancouver, Canada, July 2007, pp. 1414-1420

Power, R., Scott, D., Bouayad-Agua, N., 2003. Document Structure. *Computational linguistics*, 29:4, pp. 211-260

Rebeyrolle, J, Péry-Woodley M.-P, 1998. Repérage d'objets textuels fonctionnels pour le filtrage d'information : le cas de la définition. In: *Rencontre Internationale sur l'Extraction et le Filtrage et le Résumé Automatique, Sfax, Tunisie*, pp19-30

Shen, D., Yang, Q., Chen, Z., 2007. Noise reduction through summarization for Web-page classification. *Information Processing and Management*, volume 43, issue 6, pp. 1735-1747

Wang, G., Zhang, H., Wang, H., Yu, Y., 2007. Enhancing relation extraction by eliciting selectional constraint features from Wikipedia. In : *Proceedings of the Natural Language Processing and Information Systems Conference*, pp. 329-340