Eliciting hierarchical structures from enumerative structures for ontology learning

Mouna Kamel
IRIT, CNRS
Toulouse, France
kamel@irit.fr

Bernard Rothenburger
IRIT, CNRS
Toulouse, France
rothenbu@irit.fr

ABSTRACT

Some discourse structures such as enumerative structures have typographical, punctuational and laying out characteristics which (1) make them easily identifiable and (2) convey hierarchical relations which provide ontology fragments clues. This study will try to show how these textual objects can be exploited in order to considerably improve the process of ontology enrichment from text.

Categories and Subject Descriptors

General Terms

Algorithms, Documentation, Languages

Keywords

Ontology learning, discourse theory, document structure, enumerative structure.

ENUMERATIVE STRUCTURES

A written text is not merely a set of words or of sentences. First its semantic and rhetoric coherence must be granted by discourse relations which can be formalized through different discourse theories [5]. Second, it implements typographical, punctuation and layout means, which also contribute to identify its meaning and which can be formalized through text structure models [4]. For instance, enumeration is a feature carrying these two properties.

Enumerating consists in stating the successive elements of a same conceptual domain, these elements being hierarchically directly or indirectly linked to a classifying concept. On the textual level, this act is transcribed in a hierarchical structure, called enumerative structure (ES). The ES is made of a primer, of a list of items (called enumeration) and eventually, of a conclusion. The primer includes the classifying concept and the semantic relation that links it to the items. It introduces the list of items, an item being a co-enumerated entity which can, as afore said, be linked to the classifying concept, or eventually to another item. The conclusion, when there's one, sums up the various propositions given through the items. Also, ES structures convey hierarchical relations which provide ontology fragments clues. An ES can take several forms. It can either be written without any specific layout, or conversely be highlighted with specific typographical and/or dispositional markers.

Enumerative structure without layout

Let us take into consideration the text (T1). When facing such a text, the reader could infer that 'written language', 'sign language' and 'whistled language' are specific cases of 'non-spoken form of communication'. One can thus obtain a hierarchical structure that could well be represented by the hierarchical structure of Figure 1.

(T1) There are several non-spoken forms of communication. First, written language refers to communication in its textual form. Secondly, sign language corresponds to a gestural language. Finally, whistled language uses whistling to emulate speech.

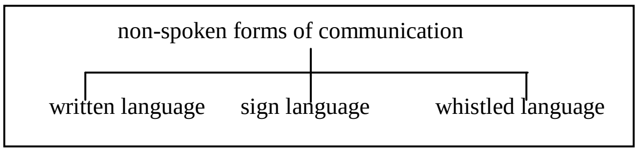

Figure 1. Hierarchical Structure from (T1)

The automatic identification of such a hierarchical structure seems out of the reach of the tools usually involved for ontology learning from text (lexico-syntactical patterns, term inclusion rule, etc). A different way to identify these inter-sentential relations is to produce a segmentation of the text and to link segments with discourse relations. These segments, sometimes named *text span* or *Elementary Discourse Unit* (*EDU*), can be contiguous or not contiguous. They are linked with subordination relation (one talks about nucleus and satellite in Rhetorical Structure Theory (RST)) or with coordinate relation (which corresponds to a multi-nuclear relation in RST) [3].

(T2) describes a possible segmentation of (T1). Each segment is annotated with square brackets and indexed with the help of a capital letter. The diagram in Figure 2 describes the rhetoric structure that corresponds to (T2) according to

K-CAP'11, June 26–29, 2011, Banff, Alberta, Canada.
ACM 978-1-4503-0396-5/11/06.

the RST. *Elaboration-Set-Member* (relation between basic information and additional information) and *List* (relation between items of the same level) are both relations of RST.

(T2) [There are several non-spoken forms of communication. **A**] [First, written language refers to communication in its textual form. **B**] [Secondly, sign language corresponds to a gestural language. **C**] [Finally, whistled language uses whistling to emulate speech. **D**]

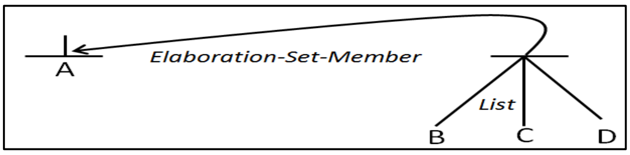

Figure 2. Rhetorical Structure of (T1)

Going from the rhetoric structure of Figure 2 to the hierarchical structure of Figure 1 is straightforward. But the steps of segmentation and representation in RST are generally carried out manually. However, these steps can be automated for some cases of SE expressed using layout.

Enumerative structure with layout

Enumerative structures we currently exploit for ontology learning from text are those which, on the one hand, are expressed with layout and on the other hand are paradigmatic, i.e. there is no dependence between items and the heads of items are syntactically equivalent. We call such structures Vertical Paradigmatic Enumerative Structure (VPES). They have the advantages of (1) being easily identifiable, (2) allowing a bijective mapping with the discourse structure it encompasses and (3) reflecting in most cases ontological discourse relations.

The translation process will need several steps: (1) the identification of the enumerative structure with layout, (2) the identification of the paradigmatic property, (3) the identification of the father and child elements together with the semantic relation that links them. Figure 3 gives an example of this mapping based on the ES equivalent to this of figure 2 but expressed with layout.

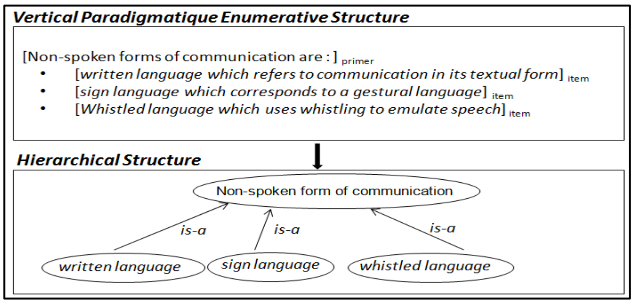
Vertical Paradigmatique Enumerative Structure
[Non-spoken forms of communication are :] primer
• [written language which refers to communication in its textual form] item
• [sign language which corresponds to a gestural language] item
• [Whistled language which uses whistling to emulate speech] item
Hierarchical Structure
Non-spoken form of communication
is-a
is-a
is-a
written language
sign language
whistled language

Figure 3. Elicitation of a VPES into a hierarchical structure

The syntactical structure of the primer enables often the identification of the father element, and identifies always the semantic relation [2]. Furthermore many studies show that authors are inclined to place the most important information at the beginning of a textual unit [1]. After having observed this phenomenon on items of enumerative structures from different corpora, we consider that the target components of the enumeration are localised at the beginning of items.

APPLICATION and PERSPECTIVES

We have enriched the OntoTopo ontology[1] by exploiting VPES from Wikipedia pages corresponding to the concepts of this ontology. In fact, Wikipedia advocates to "use the same grammatical form for all elements in a list, and do not mix the use of sentences and sentence fragments as elements".

The OntoTopo ontology has 728 concepts. We obtain 182 disambiguated pages which contain at least one VPES. From these 182 articles 434 enumerative structures including 276 VPES are extracted. Among these 276 VPES, 127 have been exploited and translated into hierarchical structures. The elicitation of these 127 VPES has provided 349 new concepts and 201 instances validated by experts involved in this project.

We intend to pursue this work according to several directions. The first one is to define learning methods for identifying the father element present in the primer of the 149 VPES not yet exploited. Another direction is to generalize our approach to other kinds of layout features or to other discourse structures such as definition. Finally, we intend to implement machine learning tools in order to improve the discovery of textual enumerative structures.

REFERENCES

[1] Ho-Dac, L-M. (2007). Exploration en corpus de la position initiale dans l'organisation du discours. Thèse de doctorat en sciences du langage. Université de Toulouse 2.

[2] Kamel M., Rothenburger B. (2010). Ontology Building Using Parallel Enumerative Structure. International Conference on Knowledge Engineering and Ontology Development (KEOD 2010), Valence, 25-28/10/2010, INSTICC - Institute for Systems and Technologies of Information, Control and Communication, p. 276-281

[3] Mann, W.C., & Thompson, S.A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. Text, 8 (3). 243-281.

[4] Power, R., Scott, D., Bouayad-Agha, N. (2003). Document structure. Computational Linguistics, 29 (2), 211-260.

[5] Wolf, F. & Gibson, E. (2006). Coherence in Natural Language: Data Structures and Applications. Cambridge, MA: MIT Press

[1] Ontology built during the GEONTO project : http://geonto.lri.fr/