Contextual Semantic Annotations: Modelling and Automatic Extraction

Noureddine Mokhtari
INRIA, 2004 route des lucioles - BP 93, FR-06902
Sophia Antipolis cedex France
Noureddine.Mokhtari@sophia.inria.fr

Olivier Corby
INRIA, 2004 route des lucioles - BP 93, FR-06902
Sophia Antipolis cedex France
Olivier.Corby@sophia.inria.fr

ABSTRACT
In this paper we propose an approach to automatically extract annotations by taking into account context in order to obtain a better representation of the document content. Our context is modelled by contextual relations built up from both the structure and the semantics of the text. Our approach requires text documents and a domain ontology as input. It automatically generates a set of contextual semantic annotations represented in RDF.

Categories and Subject Descriptors
I.2.7 Natural Language Processing, M.0 Knowledge Acquisition, M.1 Knowledge Engineering Methodologies, M.4 Knowledge Modeling.

General Terms
Design, Algorithms

Keywords
semantic annotation, annotation extraction, rhetorical relation, context, ontology.

INTRODUCTION
These last years, many works have been performed to semi-automatically extract annotations from Web resources in order to reach the semantic Web. In the field of textual semantic extraction, an important step forward has been realised through the availability of automatic natural language processing (NLP) tools. These tools are generally based on linguistic methods such as morpho-syntactic pattern matching [1] or on statistical methods such as frequency of terms co-occurrences. These works are generally limited on term extraction. Some of them enable also the extraction of relations among these terms. But in most cases, the context where these terms appear is ignored.

This observed limitation of term extraction approaches was our main motivation to propose a new approach of modelling and extracting annotations, which takes into account the context in order to give a better representation of the document content. From our point of view, the semantic annotation of a document is considered as a snapshot of its content generated by an annotator (human or program). This semantic annotation must be machine readable. Our work was carried out in the framework of the SEVENPRO European European project. The objective of SEVENPRO[1] is to develop a semantically annotated virtual environment to design products, to assist engineers in designing new products, and to allow the exploitation of both textual document semantics and 3D representations.

K-CAP'09, September 1–4, 2009, Redondo Beach, California, USA.
ACM 978-1-60558-658-8/09/09.

CONTEXTUAL SEMANTIC ANNOTATION MODELLING
We are interested in extracting contextual semantic annotations from texts. Therefore, the objects handled are of textual-type. We define *"textual object (TO)"* as a text element *(word, sentence, title, text between brackets, paragraph, section, part of a sentence…*) which conveys semantics. We define "*semantic annotation (SA)*" as the semantics (*concept(s), RDF triple(s), named graph(s)…*) conveyed by a TO. In our approach, the representation of semantics is constituted of concepts and relations associated to a domain ontology. When studying TOs, it is important to choose the right *granularity* (detail) level. The granularity level can be the "paragraphs" (resp. "sentences") and their contextual relations.

The use of any semantics is indeed tightly dependent on the context it is located in. It has to be noticed that the interpretation or the inference of a particular semantic can produce inconsistencies if the context, like *precede* or *follow*, is ignored. With regards to textual document processing, we define the context as:

"The context of a given TO is a tuple of sets <TOs,RCs> where: the TOs are the textual objects interacting with it and the RCs are the contextual relations (structural, temporal and others) implied in the different interactions".

"The context of a given SA is a tuple of sets <SAs,RCs> where: the SAs are the semantic annotations interacting with it and the RCs are the contextual relations (spatial, temporal and others) implied in the different interactions".

In contrast to the relations between *concepts* which have been proposed to represent knowledge, the *Contextual Relations* proposed here represent the relations between TOs and between SAs. Consequently, we define the contextual SAs as: "*A contextual SA is a SA with its context*".

CONTEXTUAL SEMANTIC ANNOTATION EXTRACTION PROCESS

[1] http://www.sevenpro.org/

Four main stages constitute our process: the TOs identification, the contextual relations identification, the SAs generation and the contextual semantic relations identification.

The TOs identification

The TOs identification step consists in identifying *titles, phrases, discourse markers, etc,* as well as the *arguments* of each discourse markers in the text. The GATE platform [2] has been used to identify TOs. A set of contextual relations is collected from both the *discourse markers* and other *spatial and temporal relations.* For each contextual relation of this set, a JAPE[2] rule is generated automatically to obtain their positions in the text. Other heuristics are considered and manually transformed into JAPE rules. The JAPE rules are used as transducers in the GATE pipeline.

The contextual relations identification

The contextual relations identification step requires building the text hierarchical structure: *(a)* First, the scope of the detected titles and in their hierarchy (i.e. a paragraph or a subtitle belongs to a title) are deduced; *(b)* then, the nesting among paragraphs, sentences and arguments is built by using position indicators in the text. *(c)* Finally, once the hierarchical structure of the text is built, contextual relations are deduced.

The SAs generation

The SAs generation step aims at representing the semantics of TOs within a knowledge representation formalism. The chosen formalism is RDF(S). To associate RDF triples to TOs by referring to the domain ontology, we propose to identify *classes* and *properties* in the text. Therefore, we propose to build automatically a set of JAPE rules. Indeed, the main idea is to use the value of the "*rdfs:label"* property in a RDFS schema to build JAPE rules. These rules aim at detecting the instances of *classes* and *properties* in the text. Afterwards, JAPE rules are built to detect candidate *values* of properties such as numbers in the text. Thereafter, all JAPE rules are introduced in the *pipeline* to locate *instances* of classes, *properties* and *candidate values* of properties in the text.

The generated RDF triples algorithm takes as input the TO extracted at the lowest granularity level considered (*argument*). For each TO, it identifies properties occurring in the text, and subsequently, an attempt is made to match each property with the class it is a property of and its value. If the algorithm fails to create the triple, for some properties in the text, then a larger context is sought.

The contextual semantic relations identification

The contextual semantic relations identification step of the semantic handling stage of our extraction process aims at assigning semantic roles to the discourse markers already detected. In [3][4], authors propose to automatically identify these roles (*contrast, continuation, explanations…*). However, some problems persist in complex ambiguous "discourse markers". The scope of this work is limited to the identification of discourse markers locations in the text. The role assigning of these discourse markers will be discussed in a future work.

[2] JAPE (Java Annotation Patterns Engine) is the language for expressing grammars offered by the platform GATE (an example is given in 4.2.3 section).

CONCLUSION

In this paper, we proposed an approach to model and to extract SA by taking into account the context of textual sources. The main steps of the proposed approach are summarised as follows: *i)* identification of TOs; *ii)* identification of contextual relations corresponding to TOs; *iii)* generation of semantic annotations represented by RDF triples; *iv)* identification of contextual semantic relations.

All proposed steps are automated, and a prototype is implemented to assess the various steps of this contextual extraction approach. The proposed approach has been experimented on a corpus of 2422 sentences written by the industrial partners of the European project SEVENPRO.

114 rules are thus automatically generated for the 64 classes and the 50 properties in the domain ontology that are present in the text. 80 other rules are automatically generated from the list of discourse markers as well, in order to identify contextual relations. The validation of generated RDF triples is verified manually.

The evaluation results are very satisfactory. However, the transformation of contextual relations roles into inference rules needs to be studied in more details, especially for complex discourse markers.

ACKNOWLEDGMENTS

Our thoughts are for *Rose Dieng-Kuntz* who has supervised this work and ensured its completion.

REFERENCES

[1] Aussenac-Gilles N. et Marie-Paule J. : Designing and Evaluating Patterns for Ontology Enrichment from Texts, EKAW'2006, LNAI 4248, pp. 158 – 165, Tchèque (2006)

[2] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, ACL, pp. 235-238, Budapest, Hungary, (2002)

[3] Marcu, D., Echihabi, A.: An Unsupervised Approach to Recognizing Discourse Relations, Proc. of the 40th Annual Meeting of the ACL, pp. 368-375, Philadelphia, (2002)

[4] Saito, M., Yamamoto, K., Sekine, S.: Using Phrasal Patterns to Identify Discourse Relations, Proc. of the HLTCNA Chapter of the ACL, pp. 133–136, New York, (2006