2007

Semantic processing for text mapping onto information space

Weiqiang Ou  
*University of Bolton*

Adel Elsayed  
*University of Bolton*, a.elsayed@bolton.ac.uk

Digital Commons Citation
http://digitalcommons.bolton.ac.uk/gcct_conference/8

This Conference Paper is brought to you for free and open access by the School of Games Computing and Creative Technologies at UBIR: University of Bolton Institutional Repository. It has been accepted for inclusion in Games Computing and Creative Technologies: Conference Papers by an authorized administrator of UBIR: University of Bolton Institutional Repository. For more information, please contact ubir@bolton.ac.uk.
Semantic processing for text mapping onto information space

Weiqiang Ou and Adel Elsayed
M3C Research Lab, The University of Bolton, UK
wo1ECT@bolton.ac.uk               a.eslayed@bolton.ac.uk

Abstract
This paper reports on an approach which maps documents onto an ontology-based information space in order to provide support for machine-mediated communication. First, a composite layered structure, based on a predicate-argument formalism, is proposed to provide a representation of the propositions that make up the text. We then briefly outline an approach for the automatic construction of such a structure from instructional text in the domain of physics, initially focusing on electrical circuit.

1. Introduction
The phenomenal success of the internet has resulted in the accumulation of millions of documents ready to be used, but in fact are inaccessible. We refer to these documents as digital assets. Knowledge-based digital assets codify the knowledge of an author wishing to communicate his/her thoughts to an audience. Most of the digital assets available, however, are encoded and delivered in free form language that is suitable for human consumption but only accessible to machines through metadata technology. Such a system has not been successful in developing tools to support computer-mediated communication of information.

This problem can be addressed by developing a knowledge domain which fuses the semantic content of all digital assets. The well-defined meaning is given to a document by mapping it onto a knowledge domain. The automatic procedure of such a mapping is largely accomplished through semantic processing of text.

In a general computational linguistic setting, semantic processing is responsible for transforming text into a machine-processable form, providing formal and explicit representation of semantics underlying text. Central issues in performing such a task essentially involve identifying the meaning of lexical items and structurally interpreting them in association with a specific context. Despite the significant advances that have recently been made through various knowledge-based solutions, there remain significant challenges in this realm. One of the major barriers is the lack of an agreed procedure to represent the world and domain knowledge in providing support for the process of interpreting natural language.

By limiting the scope of the domain, however, most of these difficulties can be circumvented and it becomes possible to develop formal semantics for the domain. Our research focuses on semantic processing of instructional text in the domain of physics. It is well known that instructional text possesses a set of linguistic attributes that distinguish it from other types of discourse. These differences can be manifested in terms of vocabulary used and its intended meaning, types of relationships, and the nature of the knowledge it represents. Textual material is characterized by two implicit components: objective knowledge and subjective presentation strategy. Instructional material is simply the structured and coherent presentation of these knowledge units in an attempt to achieve certain learning goals. In the context of instructional text, therefore, semantic processing can be reduced to discovering content knowledge conveyed in the text within a particular domain.

2. Predicate-argument structures
We assumed that the output of instructional text processing can be modeled as a set of atomic propositions which match factual knowledge contained in the discourse. Each proposition can be represented in the form of a predicate-argument structure, a normalized form which captures the syntactic dependency between a verb and the associated arguments. Predicate-argument structures are constructed on the basis of the meaning of lexical items and the semantic rules which are defined to build up the whole meaning recursively.

The meaning represented in a predicate-argument structure is partial, as it neglects a variety of complex factors in the linguistic domain, such as quantifications, modality, etc. Most researchers, however, believe that future NLP applications, such as machine translation or information extraction, can
hugely benefit from this simple form of semantic structure. Several techniques have been developed to identify and extract predicate-argument structure from text. For instance, Gildea and Jurafsky use statistical technologies, on the basis of hand-annotated 50,000 sentences using frame elements defined by FrameNet, to identify the semantic role of each constituent in the sentence [1]. Other works attempt to identify predicate-argument structures either on a set of predefined patterns and semantic rules [2] or on an extensive hand-written grammar, such as combinatory categorical grammar (CCG) [3].

It is to be noted, however, that a proposition derived straight from text can not be completely explicit, as it relies on the linguistic features of words. For example, the representation carry(wire, current) extracted from the sentence “the wire carries the current between two terminals in the circuit” is incomplete, due to the ambiguity of verb ‘carry’ and its failure to record ontological knowledge implicitly expressed in the sentence, e.g. wire is a conductor, conductor has free electrons which carry charges and the current is the flow of charges etc. Such a representation of proposition is insufficient to entirely capture the meaning intended by document producer.

3. Representation of propositional meaning

As an extension of the predicate-argument structure, a two-layered structure is used to meet the requirements of representing the meaning of each clause in instructional text. The first layer consists of a predicate-argument structure, including a predicate coupled with its arguments and the semantic type of the predicate. Subject, object and predicate are considered to be three building blocks of a predicate-argument structure, while each of them can be attached to their representative attributes. The ‘predicate’ is a core to this layered representation, as it, syntactically and semantically, links its arguments together. The lexical entry of verb contains information about semantic roles which expresses the nature of each of the argument involved. The ability to specify the semantic role of the arguments for verbs increases the vagueness of representation and the transparency of interface between syntax and semantics.

The second layer of the semantic representation structure is formed by identifying a subset of the predefined knowledge domain, the ontology, which is composed of concepts pointed to entities in the first layer. ‘Subject’ and ‘object’ in the first layer are linked to concepts in the ontology through lexicon-ontology mapping, while the ‘predicate’ may not find a direct relation to map to in the ontology. In such a case, a set of rules are required to derive a route which links subject and object. Thus, we end up having the text represented in two different semantic forms, each of which utilizes different sources of knowledge. It can be seen that the construction of the first layer is linguistically motivated, while the second layer is rooted in a domain knowledge structure. The clear distinction and correlation makes this layered structure a rich model of meaning representation.

4. Mapping onto information space

Driven by the proposed structure, the automatic process of obtaining propositional meaning is performed in two distinct steps. The first step involves deriving predicate-argument structures from text based on NLP technologies. Then, the resulting structures are processed according to a set of rules which aims to identify the associated portion of a predefined knowledge domain.

In our prototyped system, an ontology for electrical circuits has been developed in OWL FULL and implemented it in Protege 2000. Access to these ontologies and reasoning about them are achieved by a system that was developed on the basis of Jena [4]. In addition, OpenCCG [5], based on a CCG formalism, is chosen to provide the parsing services.

5. Conclusion

Predicate-argument structures can be viewed as a rough sketch of the entire meaning being conveyed in instructional text. This brief note suggests a two-layered structure, based on predicate-argument structures, to provide a rich and unambiguous representation of the propositional meaning of instructional text. Such semantic structures can be used to support learners in knowledge acquisition from instructional text and to support further machine-based semantic processing. The cognitive support can be achieved either directly, by users having access to and using the resulting information space, or indirectly by feeding the result of the mapping process in XML format to a cognitive-tool application that supports a specific cognitive activity e.g. reading.

6. References