User-friendly ontology authoring using a controlled language

Valentin Tablan, Tamara Polajnar, Hamish Cunningham, Kalina Bontcheva

Dept of Computer Science, University of Sheffield, Sheffield, UK {valyt, tamara, hamish, kalina}@dcs.shef.ac.uk

Abstract content

Research work in recent years in the fields of Semantic Web and Knowledge Management has started producing tools that are useful in a variety of contexts and scenarios. Many of these tools require a structuring of the information space most commonly in the form of ontologies. However, formalising knowledge in ontologies is a high initial barrier to entry for small organisations and individuals wishing to make data available to semantic knowledge technology, due to the complexity of the standards involved and the high level of experience and engineering skills required by existing ontology authoring environments.

Human language, the most natural method of communication for people, has very complex structures and a large degree of ambiguity. This makes it difficult to process automatically and machines can currently only extract a limited amount of the information therein. On the other side of the coin, formal data that is rigidly structured is easily processed by machines but hard and unnatural for people to use. The approach proposed by this paper bridges that gap by defining a controlled language which, while restricted, still feels natural to people and at the same time is simple enough and unambiguous for the machines to process.

A controlled language is a subset of a natural language which is generally designed to be less ambiguous than the complete language and to include only certain vocabulary terms and grammar rules which are relevant for a specific task. The idea of controlled languages is not new, early controlled languages can trace their roots to 1970's Caterpillar Fundamental English (CFE). The aim there was to restrict the complexity of the language used (CFE only had 850 terms) so that the text is unambiguous enough that it can reliably translated automatically into a variety of other languages. Further examples are the Caterpillar Technical English (CTE) which had 70,000 carefully chosen domain-specific terms or the KANTOO system developed at Carnegie Mellon University.

Though controlled languages can restrict the colourfulness of expression, they can be used to efficiently communicate concrete information. In most cases using a CL is an exercise in expressing information more consistently and concisely.

The controlled language proposed here is modelled to allow maximum expressibility within the smallest set of syntactic structures. The limited number of allowed syntactic sentence structures makes the language easier to learn, much easier to use than OWL, RDF, or SQL for instance. While the syntactic structure of the sentences is constrained, the vocabulary permitted is unrestricted: apart from a small number of key-phrases that are used to mark phenomena of interest, any terms can be used freely. This allows for the ontologies created to be open-domain.

The types of actions that are possible are definition of new classes, creation of hierarchies between classes, definition of object and data-type properties, creation of instances and setting of property values for instances.

The greatest advantage of this approach is that it requires essentially no training; there are no complicated user interfaces to be learnt, there are no complex formalisms to be understood. The user can simply start from a simple example which shows all the types of utterances accepted by the system and continue the ontology authoring work by re-using and modifying those examples provided. After the editing is finished, the resulting ontology can then be previewed using a simple ontology viewer implemented for this scope. Once the output has been validated, the ontology can be saved into a variety of formats including RDF-S and OWL variants.

The language analysis is carried out by an Information Extraction application based on the GATE language processing framework. It comprises some existing GATE components, i.e. the English tokeniser, part-of-speech tagger, and morphological analyser, followed by a cascade of finite-state transducers, based on GATE's JAPE pattern matching language. The role of the transducers is to search for patterns over annotations looking for constructs conforming with the controlled language. In successfully parsed sentences specific tokens are used to extract information.

The Controlled Language IE application (CLIE) employs a deterministic approach, so that each sentence can be parsed in one way only. Allowed sentences are unambiguous, so each sentence type is translated to one type of statement. If parsing fails it will result in a warning and no ontological output. In certain cases where the input is invalid, the system will try a less strict analysis mode in order to suggest how such repair may be effected.

The use of linguistic analysis allows for small variations in the surface form used to name objects (for instance the use of plurals where it feels appropriate from a linguistic point of view) without affecting the capability of the system of identifying different references for the same entity. For example the sentence `''There are animals''` will create a new ontology class with the name `'Animal'` and the sentence `''Cat is a type of animal''` will create a new class with the name `'Cat'` as a subclass of the `'Animal'` class. The `'Animal'` class is referred to in two different ways: one capitalised and in plural form and another lower case and singular. There is also support for listing items so a sentence like `''There are projects,` work packages and deliverables’’ will lead to the creation of three new classes: ‘Project’, ‘Work_Package’ and ‘Deliverable’. The names of entities are normalised – first letters are capitalised, spaces are replaced with underscores and the head word in the case of noun phrases is shown un-inflected. If this is undesirable, names can be included in single quotes which will cause them to be used as they appear in the text.

CLIE can be used in one of two modes - to create a new ontology or to add information to an existing one. Extending an existing ontology requires that names of concepts, instance, and properties in the text are first checked against those already in the ontology and only added if necessary. The domain and range restrictions of properties are also checked to ensure consistency.