Towards Automatic Construction of Domain Ontologies: Application to ISA88 and Assessment

Javier Farreres[a], Moisès Graells[b], Horacio Rodríguez[a], Antonio Espuña[b,]*,

[a] *Software Department.*
[b] *Chemical Engineering Department.*
*Escola Universitària d'Enginyeria Tècnica Industrial de Barcelona, Consorci Escola Industrial de Barcelona, Universitat Politècnica de Catalunya, BARCELONA**TECH**.*
antonio.espuna@upc.edu

Abstract

Process Systems Engineering has shown a growing interest on ontologies to develop knowledge models, organize information, and produce software accordingly. Although software tools supporting the structure of ontologies exist, developing a PSE ontology is a creative procedure to be performed by human experts from each specific domain.

This work explores the opportunities for automatic construction of domain ontologies. Specialised documentation can be selected and automatically parsed; next pattern recognition methods can be used to extract concepts and relations; finally, supervision is required to validate the automatic outcome, as well as to complete the task. The bulk of the development of an ontology is expected to result from the application of systematic procedures, thus the development time will be significantly reduced.

Automatic methods were prepared and applied to the development of an ontology for batch processing based on the ISA88 standard. Methods are described and commented, and results are discussed from the comparison with a previous ontology for the same domain manually developed.

Keywords: Standards, ISA88, Knowledge Management, Ontology, Pattern recognition.

1. Introduction

Informatics and knowledge management is increasingly recognized as a keystone for Chemical Engineering (Venkatasubramanian, 2009). Ontologies play a central role in modeling knowledge and allow users and software modules to share a consistent view of the structure of information, which enhances the reusability and scalability of software developments. Still, the creation of ontologies, particularly domain ontologies for specialized fields, requires experts to produce the ontology on the basis of their knowledge and specific language. This is a task demanding time and training: experts need to understand ontologies since knowledge cannot be systematically transferred from brains to computers. However, this may be attempted from specialized texts.

Although methods for automatic ontology construction have been reported (Hearst, 1992), the differences between domains and the corresponding specialized language indicates that a general approach is very unlikely to be produced in the near future. However, first experiences for the extraction of terminology in the biomedical domain have been reported by Vivaldi and Rodríguez (2010). Also, Küçük and Arslan (2014) have very recently presented an approach for the wind energy domain.

Hence, this paper contributes an approach to automatic / supervised ontology construction in Process Systems Engineering, based on the parsing of specialized texts, subsequent information extraction by means of a mix of textual pattern matching, linguistic analysis and grammatical parsing, and final review by human experts.

This study addresses the case of modelling batch processes according to the ISA88 standard, which is selected, in addition to its significance to Process Systems Engineering, because ISA88 provides a consistent terminology unambiguously defining a domain by means of a common model for batch control (including physical and logical models for equipment, procedures, and recipes) that has proved efficient for batch automation professionals to easily share concepts and communicate.

Process engineering has a general and growing interest in integrating information across the enterprise decision-making hierarchy. Specifically, an ontology was proposed based on ISA88 to coordinate information flows among scheduling and control decision levels (Muñoz et al. 2011). The semantic framework provided proved to allow consistent coordination of models at different time and space scales (Muñoz et al. 2012).

However, the manual development of such an ontology demanded an important effort during a long period within the development of a Ph.D. thesis (Muñoz, 2011), which also included the necessary training in ontologies and the ISA88. The ontology, limited to the Part I of the ISA88 standard, consisted of 181 concepts and 157 relations.

2. Methodology

The methodology employed is next described. The ISA88 text (Parts I to IV) was taken as a PDF document and the first task was to extract the clean text. As formatting in PDF breaks lines according to page width, line breaks had to be removed from the text, thus obtaining a text without line breaks. Next, line breaks were generated for each dot (full stop or period) in the text. In cases such as sentences containing dots ("i.e." or "etc."), line breaks were wrongly inserted, dividing sentences inadequately. Since these errors have no significant effect in the final result, they were given no further consideration.

2.1. First pattern matching stage: taxonomic relations

After all phrases were separated, the first effort was to extract the backbone relations all ontologies must have: the taxonomical relation "is a" which relates a child with its parent. Following the work by Hearst, (1992) the idea is to detect the occurrence of some patterns within each phrase. Usually, this can be carried out by simply applying plain pattern matching, but sometimes it may require some extra linguistic analysis.

This work uses a linguistic analyser, Freeling (Carreras et al, 2004), a pattern matching tool, Python (van Rossum, 1993), and a grammar parser, pyparsing (Mc Guire, 2007). In order to extract taxonomical relations, the patterns proposed by Hearst, (1992) had to be adapted to the text under study. The patterns were reduced to three productive ones.

2.1.1. Pattern 1: is a

The pattern proposed by Hearst, (1992) is NP is a NP, that is, two noun phrases separated by the "is a" construct. Whenever the construct "is a" or some variant such as "is an" or "is the" was found in a text, this sentence was processed in order to extract some taxonomic relation. For example, the phrase "The general recipe is an enterprise level recipe that serves as the basis for lower-level recipes." reveals, after a linguistic analysis, that SN("The general recipe") V("is") SN("an enterprise level recipe") that serves as the basis for lower-level recipes. Thus, the pattern SN1 V(is) SN2 allows deriving that SN2 is the parent of SN1, which in the example reflects that the concept "Enterprise level recipe" is the parent of the concept "General recipe". In order to extract this pattern and properly detect the noun phrases, a full syntactic analysis had to be performed. Pattern 1.1 resulted in 104 candidate relations.

2.1.2. Pattern 2: definitions

One variant of the previous case, and a common particular one of the ISA88 standard, is the definition of concepts. Within the ISA88 standard, some sections are dedicated to specify the definition of terms that are the keystone of the standard. These definitions take the form "id concept: definition", where "id" is the definition number identifier; the colon takes the place of the "is a" construct in previous pattern.

In this case, no linguistic analysis could be performed because the sentence was not fully constructed. Thus a shallow parsing was performed and the task was to detect the first noun after the colon. Pattern 1.2 resulted in 71 candidate relations. For example, "11 control recipe: A type of recipe which, through its execution, defines the manufacture of a single batch of a specific product." indicates that a control module is a kind of recipe, thus recipe concept is the parent of recipe control concept.

2.1.3. Pattern 3: such as

This pattern is proposed by Hearst (1992) as "such NP as NP", but in the text of the ISA88 standard no occurrence of this pattern can be observed. Alternatively, a lot of "NP such as NP" are found. Thus, the pattern was adapted to this second case.

No linguistic analysis could be performed for this case because the analyser didn't properly detect the "such as" construct to subsequently detect the two noun phrases surrounding it. A tagging process (Carreras et al. 2004) was performed in order to know the grammar category of each word and then nouns were detected before and after the "such as" construct. For example, in the phrase "Example 8: Process Management events such as allocation of equipment to a batch, creation of a control recipe, etc.", allocation of equipment to a batch and creation of a control recipe are given as examples of Process Management events, thus the "Process Management event" concept is the parent of former concepts. Pattern 1.3 resulted in 305 candidate relations.

2.1.4. Evaluation

After applying the three patterns, 480 candidate relations between 633 candidate concepts have been extracted. The 480 candidate relations have then been manually validated. 219 were found correct, 71 were found partially correct and needed manual edition, 187 were found incorrect and 3 remained undecidable. After this task, adding up the correct and partially correct results, a number of 290 different taxonomic relations were extracted relating 334 concepts.

2.2. Second pattern matching stage: meronymy relations

As the taxonomic relations are the backbone of an ontology, it was expected that all or almost all of the concepts would have been discovered at this point of the task. However, as the text of the standard is written for humans, a lot of information is implicit. So there may be non-explicit relations that wouldn't be automatically detected by the program. In this second stage, although designed for detecting new relations between existing concepts, new concepts can be discovered.

In this second stage, "part of" relations are extracted from the text. Extraction of this kind of relations has also been studied by Girju (2006). Accordingly, four patterns have been investigated and sought in the text of the ISA88 standard. For all these patterns the results of the previous stage were used. All meaningful words extracted from the first stage were used as candidate concepts for this second stage. Only those words from the previous stage are considered to detect "part of" relations, as is next described.

2.2.1. Pattern 1: part of

This pattern finds concepts in a phrase that are related by the construct "part of" within the text. For example, in the definition "Equipment unit procedure: A unit procedure that is part of equipment control." the equipment unit procedure concept is defined as a part of the equipment control concept. Pattern 2.1 resulted in 79 candidate relations.

*2.2.2. Pattern 2: includ**

This pattern tries to find concepts within a phrase that are being related by the construct "includ" within the text. Words such as "included", "includes", "including" fall in this pattern. For example, this definition "2 Formula The formula is a category of recipe information that includes process inputs, process parameters, and process outputs." explains that the formula concept includes (or is composed by) three other concepts: process inputs, process parameters, and process outputs. That is, these three concepts are parts of a formula. Pattern 2.2 resulted in 169 candidate relations.

*2.2.3. Pattern 3: contain**

This pattern finds concepts in a phrase related by the construct "contain". Words such as "contain", "contains", "contained" fall in this pattern. For example, "NOTE: An area may contain process cells, units, equipment modules, and control modules." explains that cells, units, etc. can be parts of an area. Thus, these concepts are related by a meronymy. Pattern 2.3 resulted in 193 candidate relations.

*2.2.4. Pattern 4: consist**

This pattern finds concepts within a phrase that are being related by the construct "consist" within the text. Words such as "consist of", "consists", etc. of fall in this pattern. For example, in the definition "2 Process operations: Each process stage consists of an ordered set of one or more process operations." process operation is defined as part of a process stage. Pattern 2.4 resulted in 26 candidate relations.

2.2.5. Evaluation

The three patterns generated 346 candidate relations coming from 458 different phrases relating 177 candidate concepts. Their evaluation required a harder manual job, because if the same phrase codified more than one instance of meronymy relation, only one of them were detected and the rest had to be added manually. In addition, the phrasing is much more variable than in the case of taxonomic relations, and no automatic process can expect good precision rates if no knowledge is applied to the process. After the evaluation, 254 relations were obtained (92 manually added) between 205 concepts.

Table . Comparison of the sizes of the ontologies obtained from ISA88 standard

	Manual (Muñoz, 2011)	Automatic / Supervised (this work)
Concepts / Classes	181	465
Relationships / Properties	157	544

3. Results

As given in Table 1, once the execution of these two stages was completed, a total of 465 concepts and 544 relationships had been extracted (290 taxonomic and 254 meronymic). The execution cost has been approximately

quantified in one man-month; two weeks were dedicated to program the new pattern matching algorithms, one week was required for validation. Indeed, progression was not sequential but underwent continuous improvement, and decisions were revised in regard of outcomes. Thus, engineering was the key issue and computational time was not the limiting stage.

Table 1 also gives the figures of the ontology by Muñoz (2011). Regarding the 181 concepts manually incorporated, 54 of them were detected by the automatic process after the first stage, and 38 (3 new) after the second. Therefore, 57 of 181 concepts were correctly detected, which is about 31%. It is worth noting, though, that among the 181 concepts manually incorporated there are some that do not exist in the standard (usually common sense concepts that don't need to be included in a text for humans but need to be represented in an ontology). Additionally, some concepts not existing in the standard were added for other purposes (environmental concepts, etc.). Thus, the intersection between manual and automatic outcome is much higher than this 31%.

Comparison between manual and automatic methods is difficult in quantitative terms. Suitable metrics should be proposed and used, including the effort and the quality of the ontology obtained. Table 1 provides an estimation of the completeness of the ontologies in terms of size. It’s clear that the automatic approach is producing a larger ontology with a lesser development effort (it covers all parts, I to IV, instead of only Part I).

The exhaustive, non-selective identification and extraction of concepts should be considered an advantage that reduces the chances of omitting significant concepts (false negatives) at the expense of increasing the extraction of irrelevant ones (false positives). However, the manual method is clearly prevented from including irrelevant concepts, but may fail to be wide enough. In any case, further work is required in order to carefully analyse the intersection of both ontologies (which is a manual procedure) and determine to which extent they overlap or cover different parts of the domain.

4. Conclusions

The ontology produced shows that a promising methodology has been applied that may significantly reduce the time to develop a reliable domain ontology. Tools are available that allow to be adapted and tuned efficiently in order to parse texts, perform pattern recognition from text strings, and extract concepts and relations between them. Specialized texts are also available that provide a source of knowledge in natural language and can be used to build domain ontologies.

In particular, technical standards have shown to be easier to undergo automatic knowledge extraction (purpose of clarity, definitions, glossary of terms, etc.), however, pattern recognition tools need to be adjusted since the occurrence of patterns is different from usual patterns in most texts (i.e. literature). This has been shown when using the ISA88 standard to automatically build an ontology for batch processing.

Automatic process is fast, extensive and highly productive but tends to produce the extraction of irrelevant concepts and misleading relations. Thus, a supervised procedure is required in order to manually complete the ontology. However, increasing efficiency in the development of domain ontologies may be expected from the availability and use of systematic procedures and tools for producing the core of the ontology from reliable and acknowledged technical documents.

Future work can be envisaged towards expanding this preliminary work to other standards (i.e. ANSI/ISA-95), as well as to the detection of other kinds of relations between concepts (causes).

Acknowledgements

Financial support received from the Spanish "Ministerio de Economía y Competitividad" and the European Regional Development Fund (both funding the research Projects EHMAN, DPI2009-09386 and SIGERA, DPI2012-37154-C02-01) are fully appreciated. This work has been partially funded by MINECO project SKATER (TIN2012-38584-C06-01).

References

X. Carreras, I. Chao, Ll. Padró, M. Padró, 2004, FreeLing: An Open-Source Suite of Language Analyzers. Proc. of the 4th International Conference on Language Resources and Evaluation.

R., Girju, A. Badulescu, D. Moldovan, 2006. Automatic discovery of part-whole relations. Computational Linguistics, 32, 1, 83-135.

M. Hearst, 1992, Automatic Acquisition of Hyponyms From Large Text Corpora. Proc. of Coling-92.

P. McGuire (2007). Getting started with pyparsing. O'Reilly.

E. Muñoz, 2011, Knowledge management technology for integrated decision support systems in process industries. Ph.D. Thesis. (Espuña, A.; Puigjaner, L.; supervisors).

E. Muñoz; E. Capón-García, A. Espuña, L. Puigjaner, 2012. Ontological framework for enterprise-wide integrated decision-making at operational level. Comput. Chem. Engng., 42, 11, 217-234.

E. Muñoz, E. Capón-García, M. Moreno-Benito, A. Espuña, L. Puigjaner, 2011. Scheduling and control decision-making under an integrated information environment. Comput. Chem. Engng, 355, 774-786, 2011.

van Rossum, G., 1993,An Introduction to Python for Unix/C Programmers, Proc. of the NLUUG najaarsconferentie. Dutch UNIX users group.

V. Venkatasubramanian, V., 2009, DROWNING IN DATA: Informatics and modeling challenges in a data-rich networked world. AIChE Journal, 55, 1, 2–8.

D. Küçük, Y. Arslan, Semi-automatic construction of a domain ontology for wind energy using Wikipedia articles, 2014, Original Research Article, Renewable Energy, 62, 2, 484-489.

J. Vivaldi, H. Rodríguez, 2010, Using Wikipedia for term extraction in the biomedical domain: first experiences. Procesamiento del Lenguaje Natural 45, p. 251-254.