OntoMethodus - a Methodology to Build Domain-Specific Ontologies and its Use in a System to Support the Generation of Terminographic Products

Ariani Di Felippo
NILC-FCL – Univ. Est. Paulista
P.O. Box 174 – 14.800-901,
Araraquara – SP, Brazil
+55 16 3301 6200
arianidf@uol.com.br

Sandra M. Aluísio; Leandro H. M. de Oliveira
NILC-ICMC – Univ. de São Paulo
P.O. Box 668 – 13560-970,
São Carlos – SP, Brazil
+55 16 3373 9700
{sandra,leandroh}@.icmc.usp.br

Gladis M. B. de Almeida
NILC-DL – Univ. Fed. de São Carlos
P.O. Box 676 – 13565-905,
São Carlos – SP, Brazil
+55 16 3351 8111
gladis.mba@gmail.com

ABSTRACT

Given the importance of domain ontologies for developing terminographic products, we propose a seven-step methodology – OntoMethodus – to build ontologies especially from unstructured sources. Finally, we present e-Termos, an ongoing project to develop an environment to support generation of terminographic products in Brazilian Portuguese which uses the OntoMethodus.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *language parsing and understanding*

General Terms

Languages and Theory

Keywords

Natural Language Processing; knowledge representation; ontology; Terminology; Terminography; taxonomy.

1. INTRODUCTION

Roughly speaking, domain ontologies describe the concepts that characterize a specific domain [1]. Used to systematize and model domain knowledge, they play an important role in Natural Language Processing (NLP), Electronic Commerce, Terminology, and related fields [2]. For Terminography, domain ontologies have been a starting point for the building-up of glossaries and dictionaries since they are essential for (i) identification of the concepts on which lexicalizations will be focused, (ii) elaboration of a more controlled terminology (i.e. the set of terminological units), and (iii) construction of definitions in a logical and systematized way.

Despite the importance of ontologies for the terminographic work, there is still a need to integrate tools to support all the steps of a domain ontology creation [3].

Then, we propose a *corpus*-based methodology for building domain ontologies and its integration in an environment to support the generation of terminographic products, named e-Termos [4].

This paper is organized as follows. In Section 2, we present the OntoMethodus - a seven-step methodology for building domain ontologies mainly from unstructured sources, i.e. textual *corpora*. We briefly indicate the main strategies that have been followed by computational terminologists and computer scientists in order to speed up such steps. In Section 3, we present the system e-Termos, which is followed by Conclusions and future work in Section 4.

2. THE ONTOMETHODUS

Based on the general principles of Communicative Theory of Terminology [2], we propose a methodology comprising seven stages for building domain ontologies: (a) delineation of the subject-area, (b) compilation, (c) manipulation, organization and annotation of the *corpus*, (d) terms extraction, (e) identification and representation of the ontology, (f) edition and visualization, (g) evaluation and validation. Each methodology step is detailed in this section.

2.1 Delineating the Subject-Fiel

The delineation consists of a structured view of the domain, which requires that terminologists be helped by domain experts. The delimitation has to be done based on: (a) the target audience for the terminographic product and (b) the nature of the domain itself, reflecting a specific cultural and scientific view of the reality [2].

2.2 Compiling the Specialized *Corpus*

Electronic *corpora* have strongly affected the working methods of terminologists, and brought more rigor and uniformity in terminological work. Before describing the main strategies to select texts that will compose a specialized *corpus* and to organize them, we emphasize that the term “corpus” is used here as a synonym of a set of linguistic data (pertaining to oral or written use of the language, or to both) that are (a) systematized according to settled criteria, (b) sufficiently extensive in amplitude and depth to be taken as representative of linguistic use in its totality or in some of its scopes, and (c) organized in such a manner that can be processed by computers in order to bring varied and useful results to description and analysis [Biber 1990].

In this stage, it is necessary to specify which texts should compose the *corpus*. To select “good” texts, the literature provides some criteria or guidelines. [5]: (a) originality (i.e., texts should not be translations); (b) specialized level (i.e., the difficulty of the text whether it is written for experts or general audience); (c) type (e.g., scientific, pedagogical, etc.); (d) data evaluation (i.e., authors or publisher’s reputation). With the selection criteria established, the challenge of this stage is to collect the texts. Some approaches have been proposed by using automated search engine to mining texts from Web. One of them is the BootCaT[1] toolkit, a suite of Perl programs implementing an iterative procedure to bootstrap specialized *corpora* (and terms) from the web, based on a small list of *seeds* as input.

2.3 Manipulating, Organizing and Annotating

Once compiled, the collection of texts is manipulated, organized, and annotated in order to prepare it for the automatic extraction terms process. The manipulation of the *corpus* consists of two subprocesses: data conversion and cleaning. The conversion (manual or automatic) consists of transforming documents from ps, pdf, html, doc and others to txt format. The cleaning process consists of manually cleaning or correcting the corrupted data generated by the conversion process. Once compiled from the Web and converted to txt format, the texts of the *corpus* have to be organized in a coherent way.

According to [6], there are basically two levels of *corpora* annotation. At the first, the annotation comprises the mark-up of documentation (e.g. author, publisher, edition, etc.) and structural information (e.g., chapters, sections, paragraphs, titles, lists, tables, etc). At the second level, annotation provides linguistic information about the segments in the raw text material. The added notations may include transcriptions of all sorts (from phonetic features to discourse structures): part-of-speech (PoS) and sense tagging, syntactic analysis, "named entity" identification, co-reference annotation, etc. The annotation of the *corpus* can be done with the help of natural language processing tools such as taggers, sentential segmentation tools, lemmatizers, tools to extract definitions and spellcheckers, etc

2.4 Extracting Terminological Units

Term Extraction means to (semi)automatically identify term candidates from a text *corpus*, minimizing some of the repetitive tasks associated with manual highlighting of terms. The term extraction (TE) is commonly based on three types of knowledge: (a) linguistic, (b) statistical, and (c) hybrid [7].

The proposals based on linguistic knowledge for TE basically try to identify terms capturing their (morpho)syntactic properties from tagged *corpora* [8] [9]. The proposals based on statistical knowledge commonly attempt to minimize utilization of linguistic resources, since it is difficult to compile them, and largely rely on statistical measures such as frequency, mutual information (MI), log-likelihood ratio, and Dice Coefficient, etc. Finally, the proposals based on hybrid knowledge takes into account both linguistic and statistical hints to recognize terms. Hybrid systems are usually composed by a cascade of a first linguistic analysis followed by statistical filters [8]

2.5 Identifying and Representing the Ontology

The capturing of an ontology is especially concerned with the knowledge level, i.e., independent from concerns of a particular coding language. We have distinguished between methods that apply linguistic techniques, those that apply statistics, and those that apply machine learning [10].

A paradigmatic example of taxonomic learning from *corpus* based on linguistic techniques is the Hearst’s method [11], which proposed that the presence of certain lexico-syntactic patterns indicate a particular semantic relationship between two nouns. Hearst noticed that, for example, linking two noun phrases (NPs) via the constructions *such* NP_Y *as* NP_X, or NP_X *and other* NP_Y, often implies that $\mathrm{NP_X}$ is a hyponym of $\mathrm{NP_Y}$, i.e., that $\mathrm{NP_X}$ is a kind of $\mathrm{NP_Y}$.

Within the statistical paradigm, the extraction of taxonomic relations based on clustering is the most used approach [12].

Within the machine learning paradigm, [13], for instance, first use examples of known hypernym pairs to automatically identify large numbers of useful lexico-syntactic patterns, and then combine these patterns using a supervised learning algorithm to obtain a high accuracy hypernym classifier.

Ontologies can have different degrees of formality. The two most common types are the formal and the terminological ontologies. A formal ontology is a conceptualization whose categories are distinguished by axioms and definitions. Within such paradigm, by coding, we mean explicit representation of the conceptualization in the previous stages in some formal language. This will involve committing to some meta-ontology, choosing a representation language (p.ex.: XML, RDF, and OWL) and creating the code [14].

2.6 Editing and Visualizing the Ontology

Editing tools have been developed to assist terminologists, which include suites and environments to build a new ontology from scratch or by reusing existing ontologies. Some of the more recent tools include OntoEdit, WebODE, etc. [15].

2.7 Evaluation and Validation

[Brewster et al. 2004] is an up-to-date survey of ontology evaluation approaches. The revision of literature brings four categories of evaluation: (a) those based on comparing the ontology to a “golden standard” (which may itself be an ontology); for emergent domains this approach can be a problem as the domain can suffer from having structured sources; (b) those based on using the ontology in an application and evaluating the results; (c) those involving comparisons with a source of data (e.g. a collection of documents) about the domain to be covered by the ontology, and (d) those where evaluation is done by humans who try to assess how well the ontology meets a set of predefined criteria, standards, requirements, etc.

3. THE e-TERMOS

The e-Terms (www.etermos.ufscar.br) has been designed with the aim to integrate computational tools to help the terminological researching. This environment brings together a set of linguistic and collaborative tools interconnected in a modular structure. Specifically, e-Termos can be defined as a

[1] BootCaT stands for *Bootstrapping Corpora and Terms* (http://sslmit.unibo.it/_baroni/bootcat.html).

Computer-Supported Collaborative Work (CSCW) with modules which automate the development and management of terminographic products. Broadly speaking, the e-Termos is an environment designed to allow *corpora* as input data and terminographic products as output data. In order to do that, it is largely based on the OntoMethodus. Specifically, the e-Termos is composed by six modules:

Module 0 – Automatic Compilation of the *Corpus* – responsible for extracting material for a domain *corpus* by using tools with a set of keywords (seeds) in search engines (like Google).

Module 1 – Manual *Corpus* Compilation and Analysis – responsible for manually compiling the *corpus*, to clean it and to evaluate the *corpus* quality and suitability by means of a set of natural language processing tools (e.g.: taggers, sentencial segmentation tools, lemmatizers) to extract definitions and spellcheckers.

Module 2 – Automatic extraction of term candidates from the domain *corpus* described above, using various automatic tools, from hybrid, linguistic, and statistical approaches.

Module 3 – Edition of the Ontology and Term Categorization which contains tools for creating, editing and visualizing the ontology.

Module 4 – Creation and Management of a Terminological Database, which deals with assignment of terminological information to terms, and includes the basis for establishing definitions of the terms.

Module 5 – Edition of terms and Exchange of terminographic products, which allows information assigned to the terms to be edited and the dissemination and Exchange of products available in e-Termos. It encompasses tools to export terminological data, and interfaces for editing and consulting the terminological database.

The first four modules can be used to build a domain ontology which can be saved in a format such OWL. The ontology, in this case, is the starting point for the building-up of glossaries and dictionaries. Currently, there are several database management systems for terminological data, (e.g. Corpógrafo[2]). The main differences of the e-Terms regarding to these system are: (a) the collaborative editing of terminological cards, which contains linguistic information for each terminology unit or term which is the base for the preparation of the dictionary, interacting with multiple profiles of users and being able to edit the records of terms and the definitional database, and (b) the creation of different models of terminological cards due mainly to the possibility of creating different terminographic products at the same time. Besides, e-Terms also enables the management of the definitional database and the computer-assisted definition writing.

4. FINAL REMARKS

We have described a methodology, referred to as OntoMethodus, to build domain ontologies from unstructured and structured data sources, which has been used to support the construction of the e-Termos.

5. REFERENCES

[1] Guarino, N. 1997. Understanding, Building, and Using Ontologies. International Journal of Human-Computer Studies. 46 (2-3), p. 293-310.

[2] Cabré, M. T. 1999. La terminología: representación y comunicación – elementos para una teoría de base comunicativa y outros artículos. Institut Universitari de Lingüística Aplicada, Barcelona.

[3] Corcho, O, Fernández-López, M., and Gómez-Pérez, A. 2003. Methodologies, tools and languages for building ontologies: where is their meeting point?, Data & Knowledge Engineering. 46 (1), p. 41-64.

[4] Oliveira et al, L.H.M., Aluisio, S.M., and Almeida, G.M.B. 2006. e-Termos: um Ambiente Computacional para a Geração de Produtos Terminológicos. In: X Simposio Iberoamericano de Terminologia, p. 1 – 10.

[5] L'Homme, M-C. 2004. La terminologie: principes et techniques. Les Presses de l'Université de Montreal.

[6] Ide, N., ans Romary, L. 2006. Representing Linguistic Corpora and Their Annotations. In: 5th LREC'06.

[7] Cabré, M. T., Estopà, R., and Palatresi, J. V. 2001. Automatic term detection: A review of current systems. In: D. Bourigault, C. Jacquemin, M-C L'Homme (Eds.), Recent Advances in Computational Terminology, Amsterdam & Philadelphia: John Benjamins Publishing Co., p. 53-87.

[8] Pazienza, M. T., Pennacchiotti, M., and Zanzotto, F.M. 2005. Terminology extraction: An analysis of linguistic and statistical approaches. In: Nemis 2004 Final Conference, vol. 185 (Studies in Fuzziness and Soft Computing), Athens, Greece, p. 255–279.

[9] Bernhard, D. 2006. Multilingual Term Extraction from Domain-specific Corpora Using Morphological Structure. In: 11th EACL'06, p. 171-174

[10] Gómez-Pérez, A., and Manzano-Macho. D. 2005. An overview of methods and tools for ontology learning from texts. The Knowledge Engineering Review. 19 (3), p. 187 – 212.

[11] Hearst, M. 1992 Automatic Acquisition of Hyponyms from Large Text Corpora. In: 14th Coling'92.

[12] Maedche, A., and Staab, S. 2004. Ontology learning, In: Staab, S. and Studer, R. (Eds.) Handbook on Ontologies. Berlin: Springer-Verlag, p.173–190.

[13] Snow, R., Jurafsky, D., Andrew, and Y. Ng. 2005 Learning syntactic patterns for automatic hypernym discovery. In: NIPS, vol. 17.

[14] Staab, S., and Studer, R. (Eds.). 2004. Handbook on Ontologies. Berlin, Heidelberg: Springer Verlag.

[15] Mizoguchi, R. 2004. Ontology engineering environments. In: Staab, S., Studer, R. (Eds.). Handbook on ontologies. Berlin, Heidelberg: Springer-Verlag, p. 275-298.

[2] www.linguateca.pt/corpografo/