E. Malaia

DOMAIN ACQUISITION FOR ONTOLOGICAL SEMANTICS: METHODOLOGY AND PRACTICE (DIGITAL IDENTITY MANAGEMENT DOMAIN)[1]

The present paper deals with the following methodological questions in domain acquisition for two of the static knowledge sources of the Purdue Ontological Semantics project, the ontology and the lexicon:

1) delimitation of the expanding digital identity management textual corpus with volatile vocabulary;
2) extraction of lexical items pertaining to the domain;
3) building ontological support for lexical items; introduction of necessary attributes and relations.

In order to survey the domain of digital identity management comprehensively, I subdivided it into several areas and identified major influences in each one, using them as sources for corpora in the specific subdomain.

1. Ensuring the Validity of the Corpus

From the aspect of ontology acquisition, the literature and conceptual white papers on digital identity management serve two purposes: they provide the necessary clarification to emerging concepts in the field and validate the most important ones. They are easy to spot, since values such as *anonymity*, *privacy*, *security*, and *traceability* appear across the literature in the field; while authors differ in making value judgments upon them, the concepts which contribute to the understanding of digital identity management are made clear. The DIM literature is a source of corpora, provided that we can ensure cross-validation of both the use of lexical items and concepts in the field

[1] The research was supported by ITR initiative funded by NSF grant 0428554.

through the use of multiple sources. One of the problems, which an acquirer faces in such a rapidly developing field as digital identity management, is the need to distinguish emerging concepts and cross-applicable vocabulary from the lexical items introduced ad hoc by vendors and researchers, which will not be used by anyone else.

It is important to keep track of sources of the particular corpora. Depending on the topic, the interests and goals of the various agents in the field very widely. The difference is not simply in the terminology (which can be, in some cases, quite similar), but the attention to particular aspects of transactions involved in digital identity management. Linguistically, we are interested in accurate semantic and world-view-information descriptions of all the terminology of the domain.

2. Delimiting the Corpus

For the above reasons, I have constructed the topic-source variability matrix, which deals with all aspects of digital identity management. The top row in the matrix represent the agents concerned with a particular topic (NGOs, businesses, US and international government groups, academic research). The topics (Biometrics, Psychology of DIs, technical implementation schemes, economic viability, social aspects, legal aspects) are represented by the vertical left column, and cover all aspects of digital identity management.

Based on the corpus, as discussed above, I made methodological decisions on acquisition. First of all, it was necessary to introduce new properties to the ontology in order for the entries to specify the pathways of information exchange. Secondly, it was obvious that the entire domain of «virtual world» was not represented in the ontology and had to be added. Besides, there was a need for description of hardware and software involved in digital identity management.

3. Corpus-based pre-acquisition methodology.

Before acquisition of ontology and lexicon for the domain could begin, it was necessary to create the basic structure for the ontological sub-trees and determine the lexical items which need to be added. Once the most necessary conceptual framework for the domain has been introduced in ontological concepts, the rest of the work is adding new lexical items to the domain. Also, lexical items are directly available from the corpus, as opposed to concepts, which need to be determined on the basis of their general contribution to the ontology and placed within the ontological hierarchy.

Table 1 introduces step-by-step methodology for domain pre-acquisition. The two-pronged approach enables the acquirers to ensure external validity and internal consistency of the ontology and the lexicon, and aids in faster saturation of the lexicon of a particular domain. While the topic-source subdivision is necessarily domain-specific, the two-prong methodology is applicable to ontological and lexical acquisition for any domain.

Table 1. Two-prong methodological approach to domain pre-acquisition

Top-down methodology	**Bottom-up methodology**
1. Delimit the corpus, dealing with all the aspects of DIM (topic-source matrix). Split it in two parts for validity check.	1. Run a corpus item from each topic-source combination from the matrix through the available lexicon and filter out lexemes which are not available
2. Map out an ontological tree for the most important concepts for each subdomain; establish the necessary properties for the domain overall and acquire those which are not already in the ontology.	2. Sort the lexical items as to whether they belong to Digital Identity Management domain.

3. Create ontological sub-hierarchies needed to support subdomains.	3. Acquire lexical items. Add non-domain lexical items to «IOU»/ common word-stock list (also used for running the corpus through).
4. Decide on multi-word expressions (phrasals) necessary for the vocabulary.	
5. Check for multiple meanings of available items in the lexicon, so that all the necessary word senses in the domain are represented	4. If saturation of the lexicon is not sufficient, expand the corpus.
6. Result: ontological hierarchy and lexicon for Digital Identity Management domain.	

The top-down acquisition involved, first of all, adding new properties to the ontology. The properties allow for a rigorous description of concepts in the ontology, but their number has to be limited to make future acquisition easier (not introducing limited-use concepts). The acquisition of properties is driven by both the question of grain size for the ontological description and the need for deep semantics in the description of lexical and ontological entries.

The last step in ontological acquisition was to check whether all necessary meanings of lexical items were represented in the items already in the lexicon.

In order to extract the lexical items from the corpus, I wrote a program that runs the corpus (one article at a time) through the already-existing lexicon. It also does morphological analysis, eliminating some of the morphological forms of existing words. The output of the program is a file with all the words that were not found in the main lexicon or «common words» file. The lexical items from the output file are consecutively sorted into «domain» and «non-domain» items. It is a flexible division, based on the following criteria:

1) Can the lexical item be ontologically described using the properties added for the domain of digital identity management?

2) Is it conceptually related (at least in one of its senses) to the lexical items already acquired? (e.g. *odor* was necessary for the domain of biometrics).

3) Does it appear more than once in the corpus of the domain?

4) Does it contain semantic information belonging to text meaning representation (e.g. irregular tensed verbs or some adverbs), in which case the lexical item has to be processed by the analyzer and does not have to be a part of the lexicon?

The last step of the corpus-based approach is the validity check for the domain. Optimally, one should ensure that all the lexical items occurring in the domain are accounted for in the lexicon.

The main purpose of the filtering program is to aid in lexical acquisition, and not prove that the domain is sufficiently covered. For the purposes of the present work, I rely mainly on the topic-source matrix to provide reliability for domain coverage. The other important measure of validity in lexical and ontological acquisition is the coherence of the domain in the ontology and deep semantic description of lexical items belonging to the domain.

4. Significance of proposed approach to domain acquisition.

The significance of the above research is in the development of corpus-driven acquisition methodology, which can be used further for domain acquisition. The main methodology thus far[1] has included paradigmatic approaches, such as rapid propagation approach, and «lexical rules» approach based on systematic relationships between classes of lexical entries. Corpus-driven methodology is more useful for «forming» domains with multiple emerging concepts and unstable use of lexical items, and presents a different sort of challenge, both in the choice of corpus for the purposes of acquisition and in the specifications of grain size of the entries.

[1] *Nirenburg S., Raskin V.* Ontological Semantics. Cambridge, MA: MIT press, 2004.