Automatic Knowledge Extraction and Knowledge Structuring for a National Term Bank

Tine Lassen
Copenhagen Business School,
Denmark
tla.isv@cbs.dk

Bodil Nistrup Madsen
Copenhagen Business School,
Denmark
bnm.isv@cbs.dk

Hanne Erdman Thomsen
Copenhagen Business School,
Denmark
het.isv@cbs.dk

Abstract

This paper gives an introduction to the plans and ongoing work in a project, the aim of which is to develop methods for automatic knowledge extraction and automatic construction and updating of ontologies. The project also aims at developing methods for automatic merging of terminological data from various existing sources, as well as methods for target group oriented knowledge dissemination. In this paper, we mainly focus on the plans for automatic knowledge extraction and knowledge structuring that will result in ontologies for a national term bank.

1 Introduction

If a term bank does not contain a sufficient number of terms, users will not feel encouraged to use it, and on the other hand, users will be frustrated if a term bank contains a large amount of terms with only little or poor quality information. Therefore it is necessary to use automatic procedures in order to extract and systematize information about terms, and the high quality that can be obtained by hand crafting the contents and the large volume that can be obtained by reusing terminology data from existing sources of varying quality must somehow be combined. One way of increasing the amount of terms in a term bank is to extract terms and information about terms automatically from texts. Another method is to merge terminology from different sources, such as other term banks or existing term lists. However, this approach will often lead to problems, since the term bank will typically contain many entries connected to the same term, but with varying formulation of the definitions and/or different translations. In order to clarify and distinguish the meanings of domain specific concepts, these must be described by means of characteristics and relations to other concepts, i.e. in the form of domain specific ontologies (or concept systems). On the basis of such ontologies, it is possible to develop consistent definitions that further the understanding and correct use of terms. Terminology work that includes development of ontologies is, however, a very labor-intensive task, and therefore most term banks do not include ontologies.

This paper describes our plans for automatic extraction of terms and information about terms as well as the automatic construction of ontologies on the basis of the extracted information. At present we have developed a prototype for retrieving relevant texts. We will describe this briefly in section 3.1.

Another goal of the project is to develop methods for automatic merging of terminological data from various existing sources; a problem that existing term banks have not solved adequately. The project also aims at developing methods for automatic construction of ontologies on the basis of definitions from the various data sources and methods for automatic merging of entries based on the merging of these ontologies.

Finally the project aims at developing methods for target group oriented knowledge dissemination. Many other term banks only offer restricted possibilities for setting up user specific search and presentation profiles.

As an introduction to the description of the current project we present some central concepts related to terminological ontologies.

2 Central concepts related to terminological ontologies

The backbone of terminological concept modelling is constituted by characteristics modelled by formal feature specifications, i.e. attribute-value

pairs. The use of feature specifications is subject to principles and constraints described in detail by Madsen, Thomsen, & Vikner (2004). Subdivision criteria, which have been used for many years in terminology work, were formalised by introducing dimensions and dimension specifications. A dimension of a concept is an attribute occurring in a (non-inherited) feature specification of one or more of its subordinate concepts. A dimension specification consists of a dimension and the values associated with the corresponding attribute in the feature specifications of the subordinate concepts: DIMENSION: [value1| value2| ...].

3 Subprojects

The current term bank project consists of three main subprojects: 1) Knowledge acquisition, 2) Knowledge structuring and 3) Knowledge dissemination. Figure 1 gives an overview of the project and its three subprojects as well as the processes involved. In subproject 1) Knowledge acquisition methods for a) automatic knowledge extraction and b) automatic merging and quality assurance of data are to be developed. Below, the three subprojects are briefly described.

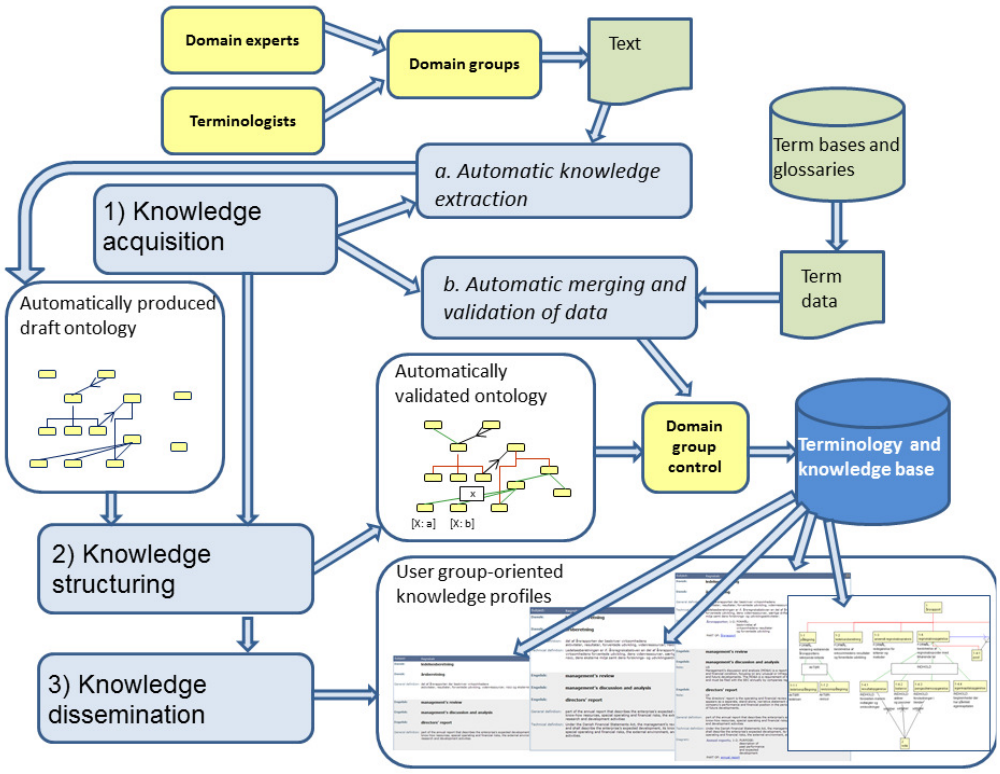

Figure 1 Outline of the project and its subprojects

3.1 Knowledge acquisition

The primary aim of the subproject 'Knowledge acquisition' is to develop new advanced models of and methods for automatic extraction of concepts and information about concepts. We develop a prototype which, on the basis of an existing domain-specific text corpus or domain texts automatically collected from the Internet, can automatically extract terms and relations and produce a draft version of a terminological ontology. The draft ontologies will contain subdivision criteria and characteristics as formal feature specifications on concepts.

One of the main ideas in this subproject is to investigate how to put together and make use of groups of domain experts, who together with terminologists in so-called domain groups (cf. figure 1) contribute to the collection of knowledge as well as to conceptual clarification.

Tools for knowledge extraction will be implemented and integrated into an interactive interface where domain experts can upload texts into a text corpus, and methods to automatically analyze these texts with respect to their (estimated) level of explicit knowledge, term density and

other LSP features (cf. e.g. Barrière, 2006 and Halskov, Braasch, Haltrup Hansen & Olsen, 2010) will be investigated.

Corpus texts will also be collected from the Internet by application of text classification algorithms. At present, in our prototype, we apply a bootstrapping algorithm, cf. BootCat (Baroni & Bernardini, 2004), where first, a small number of exemplary texts from the given domain are analyzed by applying selected statistic scores, and as a result a set of domain specific wordings or term candidates is produced. we apply co-occurrence scores, e.g. Pointwise Mutual Information (Church & Hanks, 1993) and Dice coefficient (Smadja, 1993), as well as 'termhood' scores, e.g. Log Odds Ratio (cf. e.g. Everitt, 1992) and weirdness (Ahmad et al., 1999), on n-grams, and produce a set of domain specific terms and other types of domain specific language usage that can either be the union or the intersection of the sets of term candidates produced by applying each statistic score. This set is then used as search terms, and a new collection of domain texts retrieved. The analysis and search process is iterated a number of times, until a satisfactory corpus is compiled. The definition of 'a satisfactory corpus' is still being investigated.

Another aim of this subproject is to develop methods for converting and combining terminology data from various existing sources. Two very complex types of problems exist in this process. The first type of problems that are likely to be encountered pertains to form: The data are likely to have different structures and be stored in different formats. The second type of problems pertains to content: The data may be of varying quality, and entries from the various resources may contain information about the same concept, but be associated with different sets of synonyms and with slightly varying definitions, or the other way round, have overlapping form but be associated with different concepts. Therefore, the aim of subproject 1) is also to do research in automatic ontology construction on the basis of existing term collections, and to develop methods for merging and quality assurance of term data from different sources.

3.2 Knowledge structuring

The aim of the subproject 'Knowledge structuring' is to develop methods and a prototype that may be used for automatic validation and dynamic expansion of the draft ontologies that result from the automatic knowledge extraction.

As mentioned above in section 3.1, the draft terminological ontologies will contain subdivision criteria and characteristics as formal feature specifications on concepts. This information can be used in the automatic validation of the draft ontologies: For example, if the draft ontology contains two given concepts that have been placed in a direct type relation, but where the feature specifications imply that a concept should in fact exist between them, the system can introduce a dummy concept in order to make the ontology valid. Afterwards, a domain expert must re-validate the ontology and fill in actual concepts in place of the introduced dummy concepts.

The validation process will require changes to be made in the ontology, and for this process to be performed automatically, we will develop techniques for automatic classification of concepts into ontologies with type relations based on the feature specifications that have been identified for a given concept.

Prior research distinguishes between characteristic features and conceptual relations (Madsen, Thomsen & Vikner, 2004). In the knowledge acquisition prototype, which will be developed during project subpart 1, no distinction will be made between attributes and relations per se, but all associative relations will be recorded as attribute-value pairs. For any given concept, a given characteristic feature may either be represented as a feature specification or as a relation to another concept. In a small terminology project, concepts outside the narrow domain will typically not be included in the ontology, but only exist as values of feature specifications, but if these concepts are relevant to the description of the domain, they may be included as concepts in the ontology. The project will develop new theories for distinguishing between characteristics and related concepts based on how central the values are in the given domain.

Other problems that the project will treat are multiple values and hierarchically typed values:

The knowledge acquisition prototype will potentially describe concepts with more than one (identical) relation to other concepts. However, some relations exist that can only occur once in connection with a given concept; for instance, no concept can have more than one instance of the relation HAS_LENGTH. This corresponds to the principle that a concept can have at most one value for a given attribute. Therefore, in order to facilitate ontology validation, we will develop methods for distinguishing between relations that

can only occur once, and relations that can occur several times in connection with a concept.

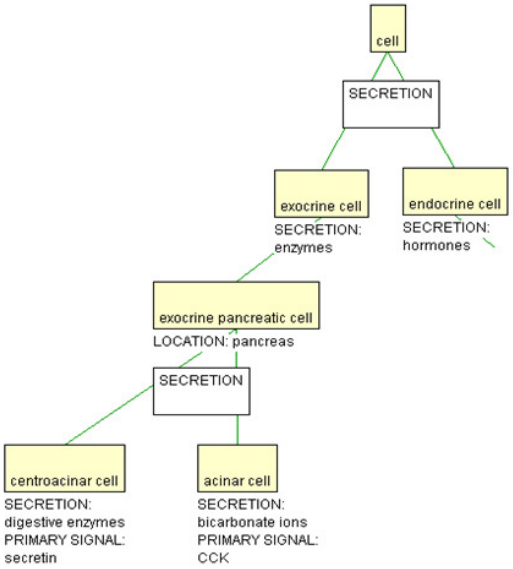

Figure 2 Excerpt of a cell ontology

In the ontology excerpt shown in figure 2, the concept *cell* is subdivided into *exocrine cell* and *endocrine cell*, based on the subdividing criterion SECRETION. The concept *centroacinar cell* inherits the feature [SECRETION:enzymes] from *exocrine cell*, but is already specified with the feature [SECRETION:digestive enzymes]. In this case, it can be argued that the value is a specialization of the inherited value, and therefore there is no conflict. To handle this, we suggest to apply a type hierarchy of values. This approach builds on the methods implemented in e.g. the Lexical Knowledge Base system (LKB) (Copestake, 1992) for use in lexical semantics.

3.3 Knowledge dissemination

The subproject 'Knowledge dissemination' will focus on presentation of data in the term bank. Traditionally, terminology and lexicography have been separate research fields with different approaches to compilation and presentation of data. However modern technology offers unlimited opportunities to meet the needs of several target groups in one database by offering the possibility of choosing between different presentations. The overall objectives of this subproject are to discuss and specify the extent to which the traditional lexicographical and terminological methods may be fruitfully combined, allowing the presentation of concepts in one single database thereby contributing added value for a defined user group, and how a combination of the two research fields may create further opportunities towards developing principles for target-group oriented knowledge transfer.

4 Conclusions

A distinctive feature of our approach includes the automatic extraction of concepts and associative relations, which can be formalised as feature specifications. The ontologies will be based on the principles for terminological ontologies as described above. No other methods or systems exist for automatic construction and consistency checking of terminological ontologies that comprise subdivision criteria and dimension specifications, which are crucial in the development of such ontologies.

References

Ahmad, K., L. Gillam and L. Tostevin. 1999. University of surrey participation in TREC8: Weirdness-indexing for logical document extrapolation and retrieval (WILDER). In: *The Eighth Text REtrieval Conference (TREC-8).*

Baroni, M. and S. Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004*.

Barrière, C. (2006). TerminoWeb: A Software Environment for Term Study in Rich Contexts. *International Conference on Terminology, Standardisation and Technology Transfer (TSTT 2006).* Beijing.

Church, K. W. and P. Hanks. 1990.Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Copestake, A. 1992. *The Representation of Lexical Semantic Information.* Doctoral dissertation, University of Sussex.

Everitt, B. 1992. *The Analysis of Contingency Tables*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability), 2nd edition.

Halskov, J., Braasch, A., Haltrup Hansen, D., and S. Olsen. 2010. Quality indicators of LSP texts – selection and measurements. How to measure the terminological usefulness of a document from a particular domain in the task of compiling an LSP corpus. *Proceedings from LREC.* Malta.

Madsen, B. N., Thomsen, H. E., and C. Vikner. 2004. Principles of a system for terminological concept modelling. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, *Vol. I*, pp. 15-18. Lisbon.

Smadja, F. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143-177.