Extracting Taxonomic Relationships from On-Line Definitional Sources Using LEXING

Judith Klavans
Columbia University
540 W. 120th Street
NY, NY, 10027
212-939-7117
klavans@cs.columbia.edu

Brian Whitman
Columbia University
540 W. 120th Street
NY, NY, 10027
212-939-7108
bwhitman@minnowmatch.com

ABSTRACT

We present a system which extracts the genus word and phrase from free-form definition text, entitled LEXING, for **Lex**ical **In**formation from **G**lossaries. The extractions are used to build automatically a lexical knowledge base from on-line domain specific glossary sources. We combine statistical and semantic processes to extract these terms, and demonstrate that this combination allows us to predict the genus even in difficult situations such as empty head definitions or verb definitions. We also discuss the use of 'linking prepositions' for use in skipping past empty head genus phrases. This system is part of a project to extract and structure ontological information for energy-related glossary information.

Categories and Subject Descriptors

[**Representation Issues**]: Ontologies and taxonomies.

Keywords

Ontologies, glossaries, definitions, lexical knowledge bases (LKB), information retrieval, natural language processing.

1. USING DEFINITIONS TO BUILD TAXONOMIES FOR DIGITAL COLLECTIONS

The research we present is part of a larger project to improve access to a set of heterogeneous databases in the Energy Data Collection project of the Digital Government Research Center. As part of this project, a large domain specific ontology serves as the information broker to improve access across databases. The domain specific glossary items will be merged with a larger ontology. We use glossary definitions since they are by their very nature domain specific, that is, they give information only pertinent to the glossary's domain. For example, the following definition appears in our data set:

> **Motor Gasoline Blending Components:** Naphthas (e.g., straight-run gasoline, alkylate, reformate, benzene, toluene, xylene) used for blending or compounding into finished motor gasoline. These components include reformulated gasoline blendstock f or *{sic}* oxygenate blending (RBOB) but exclude oxygenates (alcohols, ethers), butane, and pentanes plus. *Note:* Oxygenates are reported as individual components and are included in the total for other hydrocarbons, hydrogens, and oxygenates.

Figure 1 - Sample Glossary Definition

Compare this relatively unstructured entry with a typical dictionary definition which has rich internal structure e.g. pronunciation, etymology, sense numbers, etc. In the dictionary analysis projects in [1], over fifty of such fields are identified. To achieve our goals for the EDC project, we must:

1. Identify the definitional material from a web page or other online source (*identification*) (see [2])
2. Parse the definition for its most salient properties and features (*LKB generation, genus finding*)
3. Incorporate the structured information into a larger ontology, including linking and merging definitions from different agencies and sources.

2. GENUS TERM AND PHRASE FINDING

Our model of building a lexical knowledge base (LKB) from machine readable dictionaries (MRDs) is influenced by the work done in [3] and [4], in which the authors propose a hierarchical structure to represent the complex information found in MRDs. Richly embedded structures containing a head word with subordinate information including cross-references are derived from MRDs. The structured LKB can then be queried as a database and information which was previously inaccessible becomes available. The key feature of this research is that definitions (e.g. car: a vehicle with four wheels) consist of a genus term (vehicle), defined as the main noun or noun phrase which captures an is-a type relation, and differentia (with four wheels), which details how the defined term differs from related terms (e.g. motorcycle: a vehicle with two wheels).

The method LEXING uses to determine genus phrase and terms uses knowledge gained from the part-of-speech tagging and noun-phrase chunking [5] components. We have developed a grammar of phrase identification from a manual study of various definitional sources, and have implemented a evaluation metric for comparing our system's results against a manually tagged set of 500 glossary definitions from 5 different sources.

Definition → (Head Term:) (Definition Text)
Definition Text → (Genus Phrase) (Remainder)
Remainder → Text
Genus Phrase → NP (of)? (Genus Phrase)
Genus Term → (last noun of first GP NP)

Figure 2 - Genus Phrase and Term Grammar

Since each domain could use domain specific semantic separations, we also introduce the notion of *automatically derived* *semantic attributes* that are inferred simply from their frequency in the text. The LEXING system identifies separators, such as *having a, used for,* or *containing a*, are "cue phrases" that identify the next clause as suitable for semantic chunking. As an example, the phrase *having a* was not in our original list of manually-derived separators, but after running a bigram analysis, we discovered its frequency and importance.

3. RESULTS OF LEXING EVALUATION

We have performed two evaluations on LEXING output:

1. Definition Content Analysis: we ran our system on various definitional sources to determine if our ideas of content were correct, and which fields are frequent.
2. LEXING Accuracy: we evaluated the genus term identification algorithm against a "gold standard".

For evaluation of our representation of definition content, the semantic separator components were tested using both unedited and edited definitions. Our main definition sets came from two government agencies: the Energy Information Administration (data on energy sources such as gasoline or coal), the Environmental Protection Agency (glossaries on environmental concerns); we also tested over heart-disease related definitions from definition extraction work done in [5]. Inputs ranged from web pages to flat ASCII documents.

Table 1. Definitional Content Analysis

	Terms	Genus Phrases	Prop-erties	Quant-ifiers	Includes exclude
EIA Edited	19	18 **95%**	15 **79%**	7 **37%**	2 **11%**
EIA Web	127	121 **95%**	38 **30%**	50 **39%**	9 **7%**
EPA Web	1054	1029 **98%**	56 **5%**	24 **2%**	75 **7%**
Medical Auto	90	83 **92%**	0 **0%**	0 **0%**	0 **0%**

The results in Table 1 show that definitional content hinges largely on the genus phrase and word. Semantic properties are found in well-edited glossaries (such as EIA edited). Note that the results above do not indicate LEXING's performance. Rather, this evaluation indicates the profile of source definitions in terms of complexity.

For the second evaluation to determine the accuracy of LEXING, we manually tagged 500 definitions, 100 each from 5 domains. (Civil Engineering, Computer Terms, Biomedical Information, General Medical Information, and Energy Information) in order to establish a measurement standard or "gold standard". To compute scoring for genus term identification, a match was defined as both the human tagger and the computer choosing the same genus term. We then computed the accuracy over each definitional set. To compute scoring for genus term identification, a match was defined as both the human tagger and the computer choosing the same genus term.

Table 2. Genus Term Finding Results

Domain:	Civil Eng	Comp.	Biomedical	Medical	Energy
Genus Term	93/99 **94%**	81/101 **80%**	93/102 **92%**	100/103 **97%**	85/102 **83%**

Through the first experiment outlined in the above section, we continually isolated and fixed errors related to the semantic attributes and separators. Although some of our definitional sets did not contain as much extractable content as we would have liked, the results on the edited sets were more fruitful. Through this feedback mechanism (run with known set, check results, and fix) we also improved our genus term extraction to achieve the strong results shown in the second experiment.

4. CONCLUSION

We show that a combination of semantic knowledge (the semantic separators) and statistical methods (the automatically-derived semantic attributes) can together provide a methodology of deriving lexical knowledge bases and specifically genus terms, from free-form glossary sources. We draw on the previous machine-readable dictionary to lexical knowledge base research. What is novel is our methods of processing flat unstructured input of on-line glossaries and the processing of automatically-identified multiple glossary sets from different agencies. Our ultimate goal is to merge this information into a large on-line ontology.

5. REFERENCES

[1] Byrd, R.J., B.K. Boguraev, J.L. Klavans and M.S. Neff. From Structural Analysis of Lexical Resources to Semantics in a Lexical Knowledge Base. U. Zernik (eds.) Proceedings of the First International Workshop on Lexical Acquisition, Detroit, Michigian, 1989.

[2] Whitman, Brian. Automatically Determining Definitional Content from Web Documents. Technical Report, in progress. Columbia University, 2000.

[3] Neff, Mary and Bran Boguraev. Dictionaries, dictionary grammars and dictionary entry parsing. Proceedings of the Twenty-seventh Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 1989.

[4] Klavans, J., Chodorow, M., & Wacholder, N. From dictionary to knowledge base via taxonomy. Proceedings of the 6th Annual Conference of the UW Centre for the New Oxford English Dictionary. Waterloo, Ontario, pp. 110-132, 1990.

[5] Evans, D., Klavans, J. and Wacholder, N. (2000) *Document Processing with LinkIT.* Recherche d'Informations Assistée par Ordinateur (Content-Based Multimedia Information Access.) Paris, France, pp. 1336-1345.

[6] Klavans, J.L., Muresan S. DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-Line Text. Proceedings of AMIA Symposium 2000; p. 1096, 2000.Evans, D., Klavans, J. and Wacholder, N. (2000)