Building a Terminological Database from Heterogeneous Definitional Sources

Smaranda Muresan
Columbia University
New York, NY
smara@cs.columbia.edu

Samuel D. Popper
Columbia University
New York, NY
sp2014@cs.columbia.edu

Peter T. Davis
Columbia University
New York, NY
ptd7@cs.columbia.edu

Judith L. Klavans
Columbia University
New York, NY
klavans@cs.columbia.edu

Abstract

An obstacle to understanding results across heterogeneous databases is the ability to determine conceptual connections between differing terminologies. In this paper, we present the two step approach which we have used to build a terminological database in order to address this issue. First we automatically built a heterogeneous collection of terms and definitions from two types of dynamic sources: 1) glossaries automatically identified from 147 government web sites and 2) definitions extracted from 600 unstructured articles. After storing terms and their definitions, we semantically analyzed the definitions to store the terminological knowledge in a relational database. Currently the database contains 12,780 definitions of 8,431 terms.

1 Motivation for a Terminological Database

Due in part to the rapid growth of the Internet, individuals and researchers have unprecedented access to data from many sources. Failure to properly understand the concepts behind these data sets may lead to erroneous conclusions in data analysis. We have built an automated system that can discover and make available concepts, definitions and inter-definitional relationships from several Internet sources by building a terminological database.

In building this database, we reflected two desiderata for terminological resources: 1) to enable representation of the ongoing evolution of language since new words and senses continually appear in language, and 2) to allow users to explore the richness of terminological knowledge (e.g how terms relate to each other, what are their semantic properties and attributes). Our goals are to provide easy and efficient access for users of multiple information sources across government sites.

Towards meeting these desiderata, as a first step, we automatically built a heterogeneous collection of terms and definitions from dynamic sources. As a second step , we semantically analyzed these definitions in order to identify relations among terms and their attributes. We then built a relational database. The architecture of the system is shown in Figure 1.

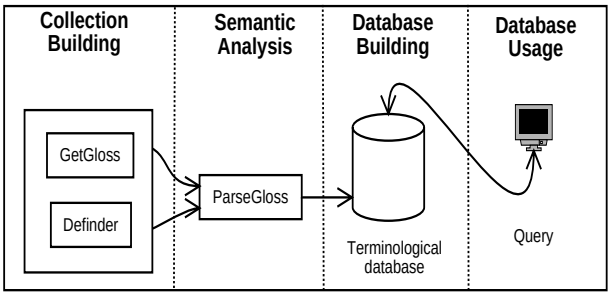

Figure 1: Overall architecture of the system

2 Building the Heterogeneous Collection of Terms and Definitions

In order to build a dynamic terminological database, a key step is to automatically identify terms and their definitions from dynamic sources. Figure 1 shows two modules: 1) `GetGloss` for the identification of glossaries across government websites and 2) `Definder` for the extraction of terms and their associated definitions from unstructured on-line articles or other documents.

GetGloss. An officially published glossary reflects the input of highly-trained specialists in a given area, and thus constitutes a high-quality source of information. However, the task of automatically identifying glossaries from websites poses sev-

eral challenges: 1) glossaries can constitute small deeply embedded parts of a web page; 2) there is no standard HTML tag formatting for marking (term,definition) pairs; 3) a web page can contain what might appear to be (term,definition) pairs, but are not (e.g. lists of modules in system). `GetGloss` is a system for extracting glossaries embedded in web pages (Klavans et al., 2002). A generic crawler downloads pages from domains specified to it, and passes these pages to a two stage Analyzer. First, the Identification component finds lists of (term,definition) candidates in a page using keyword and rule-based algorithms. Second, the Categorizer component filters out the false candidates.

Definder. Although useful, on-line glossaries are static, incomplete, and often geared towards one domain. To supplement `GetGloss`, we have developed a system to extract definitions from unstructured on-line articles to automatically build a dynamic dictionary from text. We first developed a robust rule-based system to extract the most reliable definitional patterns, using both shallow and deep parsing. The system is described in (Klavans and Muresan, 2001). In current research, the results of this system are used as seed tuples *(term, definition)* in a bootstrapping algorithm similar to the one in (Riloff and Jones, 1999).

3 Building the Terminological Database

These two techniques yield unrelated lists of terms and definitions from multiple sources, with no coherent way to represent or use the results. The purpose of constructing a unified database was to address this problem. Our approach has two steps: 1) semantic analysis of definitions using the `ParseGloss` system, and 2) building a relational database. In this way, we move towards a more efficient way to handle the amount of data from different sources, multiple definitions for the same term that can be inconsistent with one another, and need for dynamically updating the collection. Ultimately, semantic analysis will provide semantic consistency, and a relational database provides efficient storage, fast access to data, with suitable query and updating mechanisms.

ParseGloss. Semantically parsing definitions extracted from heterogeneous sources is a challenging task. The `ParseGloss` module (Klavans et al., 2002) involves an initial shallow syntactic parse of the definition. Rule-based and statistical methods are then used to derive attributes and semantic relationships among concepts (*hypernym* which is the genus-phrase, *synonym*, *purpose*, *modifier*, etc.). This semantic analysis provided by `ParseGloss` transforms the definitions to a more structured representation that can be encoded in a relational database.

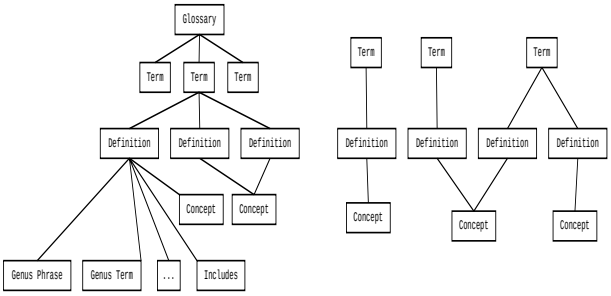

Figure 2: Two Views of the Database: Hierarchical (left) and Flat Concept-oriented (right)

The database has two main logical views as shown in Figure 2. The first is a *hierarchical structure*, based on the origin of the definitions. Each glossary contains multiple terms. Each term may contain several definitions. A definition will have associated with it a concept. A given concept may have several definitions, that comes from different sources. Another view is a *flat concept-oriented view*. The database can be considered to simply hold a series of definitions for many concepts, along with information about each concept and definition. There is a many-to-many relation between terms and concepts (expressing the notion of synonymy and polysemy). Thus an arbitrarily complex graph may emerge instead of a forest of hierarchical trees. These multiple views of the database allow expressive query facilities for the user.

The database can also encode an arbitrarily complex amount of data for a definition. A relations table is provided to relate any concept to any other concepts. An item at any level may have a property associated with it, whose semantics is defined by the particular application that uses it. The database was designed to allow for high degrees of normalization. We use XML as data transport between the defini-

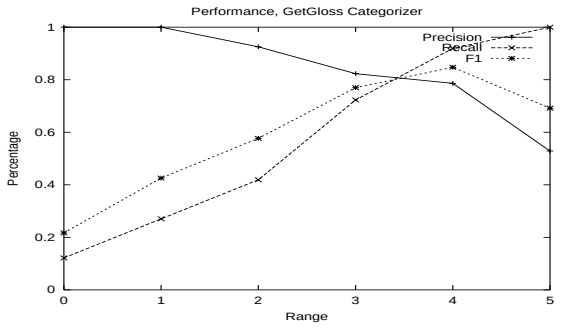

Figure 3: Precision, Recall, and F_1 for the GetGloss Categorizer for different score ranges.

Definition	Source
Redness, swelling, heat and pain resulting from injury to tissue (parts of the body underneath the skin). Also known as swelling.	GetGloss
A characteristic reaction of tissues to disease or injury; it is marked by four signs: swelling, redness, heat, and pain.	GetGloss
The body 's response to irritation , by releasing chemicals that attack germs and tissues and also repair the damage done .	Definder
A local reaction to irritation , injury , or infection characterized by pain , swelling , redness , and occasional loss of function .	Definder

Table 1: Definitions of the term *inflamation*

tion extraction modules, semantic analysis component and the database.

4 Evaluation and Experiments

4.1 Collection building

GetGloss. Since `GetGloss` is the input to a semantic analysis component, it is important to have high precision. We have modified the Categorizer component such that each candidate glossary is assigned a score through a linear combination of weighted features. If the score is below a threshold, it is considered to be a true glossary. We have tested the effect of different thresholds on precision and recall on a set of 300 manually categorized candidate glossaries randomly chosen from a set of 2400 candidates extracted by the `GetGloss` Identification component. Figure 3 shows the effect of different score ranges on precision and recall (Range 0 represents scores below -100, range 1 is scores between -100 and -51, and so on). As can be seen, between range 3 and 4 we obtained 82.12% precision and recall. Setting a lower threshold can increase the precision to above 90%, but recall decreases.

Definder. In order to thoroughly evaluate this system we focused on several dimensions: 1) performance of the definition extraction algorithm in terms of precision and recall measured against human performance; 2) quality of the generated dictionary definitions as judged both by non-specialists and by medical specialists; 3) coverage of on-line dictionaries. For the first evaluation, the rule-based version achieved 86.95% precision and 75.47% recall (Klavans and Muresan, 2001), while the hybrid method with bootstrapping achieved better recall, 84.60%, with an insignificant decrease in precision, 86.27%. The second evaluation showed that `Definder`'s definitions were accurate and complete as judged by specialist users, and useful and readable by non-specialists. The third evaluation showed that existing resources are far from being complete, and `Definder`'s output is useful to enhance these resources (Klavans and Muresan, 2001).

4.2 Semantic Analysis

To measure the accuracy of `ParseGloss`, a corpus of 100 terms and their definitions was randomly collected from a varied sample of glossaries. In both paper-based and web-based settings, human subjects were then asked to identify the *most important word or phrase* of the definition, and to list additional properties of the definition. For each additional property, subjects were asked to quantify the relationship to the term. The responses are collected to create a gold standard corpus of definitions together with their semantic analysis. We are in the process analyzing results to benchmark the current `ParseGloss` implementation, and to evaluate further improvements to the system (Klavans et al., 2003).

4.3 Database building

We performed a quantitative evaluation of the knowledge contained in the database. There are currently 12,780 definitions and 8,431 terms represented in the database from 147 government web sites and 600 articles. An example of one term with many definitions from several sources (different glossaries in case of `GetGloss`, different arti-

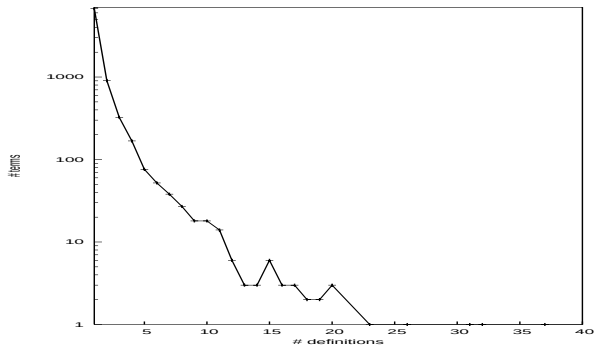

Figure 4: Distribution of the number of terms given their number of associated definitions

cles in case of `Definder`, or coming from both) is given in Table 1. Figure 4 shows the distribution of the number of terms based on their associated number of definitions. Of the 8,431 terms, 1,680 have 2 or more definitions (e.g 903 terms with 2 definitions, 76 terms with 5 definitions, 3 terms with 20 definitions). There are terms with up to 40 definitions. Also the database representation allows for query based on semantic relations identified by `ParseGloss` (e.g all terms that have the same hypernym, crossreferenced terms, and so on.)

5 Related Work

To our knowledge, collecting heterogeneous glossaries across government websites has never previously been performed. However since we see glossary identification as a categorization task, techniques used in text categorization are relevant to this research (Joachims, 1998). The extraction of definitions from unstructured on-line text is also a new problem. The closest relevant research is the question-answering task of for definitional questions (Prager et al., 2001). Our work on semantic analysis of definition is related to the work on parsing Machine Readable Dictionaries (Richardson et al., 1998). However the nature of our heterogeneous sources poses additional challenges as compared with dictionary definitions since publishers generally impose structural constraints on entries. Relevant complementary research is the work on ontology learning (Staab et al., 2001). There is also current effort on deciding the best representational models for linguistic databases that is relevant to our work (Bird et al., 2001).

6 Conclusions and Future Work

In this paper we present a step towards solving the heterogeneous terminology problem. We described a system for building a terminological database in two steps: 1) automatically building a heterogeneous collection of terms and definitions from dynamic sources and 2) semantically analyzing the definitions and building a relational database. We plan to extend the work on our semantic analysis module to be able to merge several definitions of the same term that comes from different sources, and to qualitatively evaluate the database with users.

7 Acknowledgements

This research was funded by the National Science Foundation Digital Government program, award #0091533 “Digital Government Research Center (DGRC): Bringing Complex Data to Users”.

References

S. Bird, P. Buneman, and M. Liberman. 2001. IRCS workshop on linguistic databases.

T. Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML’98*.

J.L. Klavans and S. Muresan. 2001. Evaluation of DEFINDER: A system to mine definitions from consumer-oriented medical text. In *Proceedings of The First ACM+IEEE JCDL 2001*.

J.L. Klavans, P. Davis, and S. Popper. 2002. Building large ontologies using web crawling and glossary analysis techniques. In *Proceedings of dg.o2002*.

J.L. Klavans, S. Popper, P. Sommer, W. Bourne, , and L. Tilmanis. 2003. Evaluating semantic parsing techniques using human agreement. *paper in progress*.

J. Prager, D. Radev, and K. Czuba. 2001. Answering what-is questions by virtual annotation. In *Proceedings of HLT’01*.

S. Richardson, W. Dolan, and L. Vanderwende. 1998. Mindnet: acquiring and structuring semantic information from text. In *Proceedings of COLING ’98*.

E. Riloff and R. Jones. 1999. Learning dictionaries for information extraction by multi-level boots trapping. In *Proceedings of AAAI’99*.

S. Staab, A. Maedche, C. Nedellec, and E. Hovy. 2001. The second workshop on ontology learning OL-2001.