Tackling the Internet Glossary Glut: Automatic Extraction and Evaluation of Genus Phrases

Judith L. Klavans
Center for Research on Information Access
Columbia University
212-854-7443
klavans@cs.columbia.edu

Samuel Popper
Department of Computer Science
Columbia University
212-854-7443
sp2014@columbia.edu

Rebecca Passonneau
Department of Computer Science
Columbia University
212-939-7112
becky@cs.columbia.edu

ABSTRACT

This paper addresses the problem of developing methods to be used in the identification and extraction of meaningful semantic components from large online glossaries. We present two sets of results. First, we report on the algorithm, ParseGloss, which was used to analyze definitions, and extract the main concept, or genus phrase. We ran the system on over 12,000 online glossary entries. Second, we present a method to evaluate our results, using human judgments on a collection of definitions from six different sources. This paper discusses our approach to the evaluation process, since the creation of a standard for evaluation is in itself a contribution to the field. The methods we have developed have required addressing the significant challenges of abstracting a single gold standard from multiple naive, human judgments on a highly subjective task. Once the method for creating the standard was developed, we then established the gold standard data. We report on our performance in running ParseGloss over this controlled collection of definitions. Our first set of results presents precision and recall on system performance. Our second results are presented in terms of techniques for determining agreement between human subjects. Success in the ParseGloss algorithm will contribute to the automatic creation of ontologies.

Categories and Subject Descriptors

A.2 [**Reference** (glossaries)]; E.5 [**Files**]: sorting/searching; E.2 [**Data Storage Representations**]: linked representations; H.1.2 [**User/Machine Systems**]: Human information processing

General Terms

Your general terms must be any of the following 16 designated terms: Measurement, Experimentation.

Keywords

Terminologies; Ontologies; Parsing; Natural Language Processing. Glossary analysis, computational linguistics, parsing, natural language processing, terminologies, conceptual analysis, evaluation methods.

Semantic Web Workshop, SIGIR 2003, July 28-Aug 1, 2003, Toronto, Canada.

1. EXTRACTING INFORMATION FROM GLOSSARIES

1.1 Background

In theory, the concept of a Semantic Web entails a reliable and predictable semantic theory which is coherent, complete and consistent. However, the reality is far from this goal. Indeed, as noted in many w3c.org conferences, the Semantic Web is a vision for the future of the Web in which information is given explicit meaning. In principle, this would then make it easier for machines to automatically process and integrate information available on the Web, across vocabularies and domains. This paper addresses the reality of addressing this vision, with specific focus on existing content available in on-line glossaries.

The Web contains a vast array of glossaries, some of which occur as independent glossary-only pages, but others of which are embedded in non-glossary documents. Glossaries are a rich source of semantic information, often in semi-structured form. Our goal is to create a conceptual database of these terminologies in order to enable cross-domain navigation via concepts. In previous work, we reported on the GlossIT system, which enables the harvesting of glossaries from the web, combined with a text data mining component, in order to create a terminological database. We addressed the issues of:

1. harvesting and identifying of glossaries in webpages as a categorization task, using hybrid rule-based and machine learning methods in a data collection component called GetGloss [16];
2. parsing of web pages with multiple formats into headword, definition tuples in order to load into a database [16];
3. adding a text data mining component, Definder [15,21], to add definitions and headwords identified from free-form text;
4. parsing individual definitions into conceptual components with the ParseGloss system, including recognition of genus/species and other semantically based concepts, such as includes(x), excludes(x), related-to(x).

The architecture of the entire GlossIT system is shown in Figure 1. The ParseGloss module analyzes glossary entries, whatever the source, and loads the results into a relational database, as shown in Figure 2.

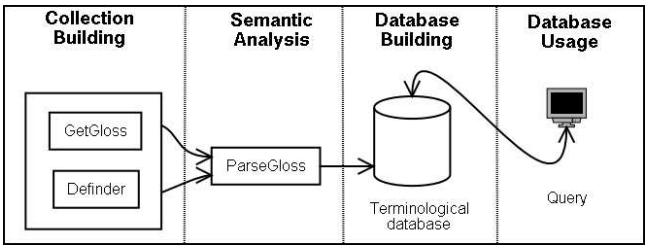

Figure 1: The GlossIt System

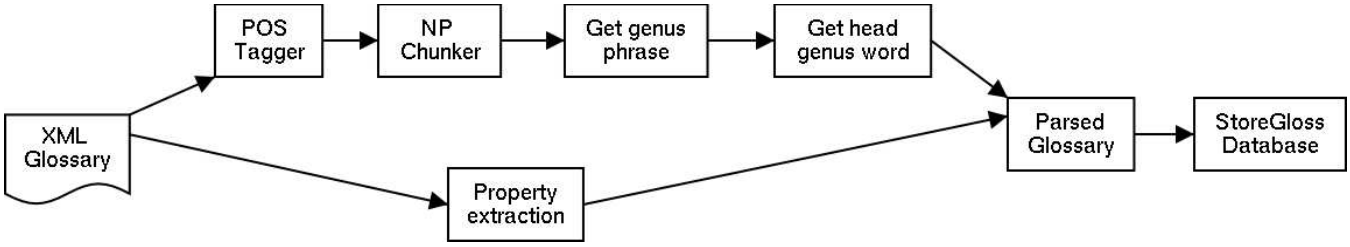

Figure 2: The ParseGloss System

From these sources, we have assembled a database of about 8,000 terms, with 12,500 definitions from 2,200 different sources. Over 6,500 terms come from 900 URLs, and 3,500 terms come from 1,350 free-text articles. Note that the sum of the number of terms is greater than the reported total. This is because of overlapping terms that occur in both web pages and free-text articles. This demonstrates that the free-text articles usefully augment the web pages, and also that they address some of the same topics as the web glossaries.

Because of the scope of the dataset, manual creation of the database would not be feasible. Nor is it desirable, for by the time any one collection has been created, indexed, and made available, new terms will have been created, and categories will have changed.

This paper reports on new results on evaluation of the ParseGloss component, and the subsequent improvements we suggest.

1.2 Problem

Challenges in this task are known to be difficult since definitions can be complex and ambiguous. Our goal is to permit more flexible access to multiply defined terms from heterogeneous databases from different government sites, and to permit the association of related terms via genus. We have initially focused on large government websites as part of a Digital Government project on unified access to heterogeneous distributed databases, although our techniques are fully generalizable across domains.

The automated analysis of terminology addresses the growing problem of information exchange, by creating links across disparate terminologies. The explosive growth of information available has presented new opportunities for researchers: in addition to the information now easily available from a given agency, data from several different agencies can be collected to provide new insights. For example, auto emissions data from the Environmental Protection Agency (EPA) might be correlated with data from the National Oceanic and Atmospheric Association (NOAA) regarding long-range weather trends. In addition, the same terms might be defined differently depending on a variety of factors, including the source agency, as discussed in section 3.

Any attempt to relate definitions from distinct web sources will encounter several problems, both in format and in content. From a structural perspective, each source may have a different layout and text representation which is an issue for collection building. Mapping into relational databases and XML will address this issue, but a more pressing problem is that of relating terms and their definitions across heterogeneous document sources. These problems of semantic content are more subtle and difficult to address. They include ambiguity, uncertainly, and incomplete information; each of these cases is discussed throughout this paper, but consider a simple case of incompleteness, and how this might affect an automatic method for extracting correctly referring and accurate commonalities. For the EPA example, does the data on auto emissions include only consumer vehicles, or does it average all vehicles on the road? Since there are a tremendous number of data-collecting organizations (federal, state, local, and non-government organizations (NGOs)) which change over time, a resource that can automatically assemble terminology from distinct agencies and highlight the commonalities and differences would serve the dual purposes of making the terminology more accessible and keeping abreast of changes.

We propose a solution based on collecting and cataloging the glossaries on government web sites. When an agency creates a glossary, it tends to put a significant amount of energy into the process, resulting in a set of terms and definitions that is official and highly reliable. As shown in Figure 1, we have built a system, GlossIT that locates candidate document sources made available on government web sites, then extracts these glossaries so as to create a single, heterogeneous super-glossary of inter-agency terminology. GlossIT contains logic to find cases where there may be short lists of glossary entires, embedded in longer files, such as book chapters or other documents. Once the glossaries are obtained and analyzed, they are loaded into a database designed to facilitate creation of relationships among term/definition pairs, and to provide a uniform query environment for querying and browsing the super-glossary.

What distinguishes this paper is that (1) we process sources from a wide range of domains, (2) we confront the breadth of unstructured styles found on the web, and (3) we extract information from glossaries.

2. RELATED WORK

The review of ontology generation techniques in Ding and Foo 2002 shows that, of the six projects reviewed, source data is found in free text, definitions, controlled vocabularies and thesauri. However, to our knowledge, none of these approaches performs the generalized web-harvesting for glossaries as in GetGloss. At the same time, what is common across approaches is the need to address the challenge of the variety of formats, styles, and domains that ParseGloss needs to handle. In particular, the complexity of definitions, shown in Section 3, in the government websites we have crawled exhibits semantic issues pertaining to the genus term, when it is present at all, that present problems for reasoning over the terminological database, such as ambiguity or vagueness. Collecting and structuring terminologies into databases has been included in many independent ontology-related projects from several domains, including WordNet [20] UMLS (http://www.nlm.nih.gov/research/umls/), and AAT (http://www.getty.eduresearch/tools/vocabulary/aat/). The Digital Government Research Center outputs to an interchange format used with the ontology created by [11].

Unlike many of the related research papers on learning and generating ontologies [18], this paper focuses on preliminary data harvesting required in order to explore methods of automatically building ontologies from web resources. Other researchers (e.g. [6]) have addressed the issue of collecting conceptual information from domain-specific text. The output of our research could be used for task such as those discussed in [17].

Another area of related research involves parsing dictionary definitions. While parsing definitions for conceptual information has been attempted over different dictionaries, including the Longman Dictionary of Contemporary English (LDOCE) [22], , the COBUILD Dictionary [23] and Webster's Seventh Dictionary (1982), these efforts have been focused primarily on dictionary definitions [3, 13, 14, 10]. In contrast, the ParseGloss effort has been aimed at heterogeneous glossaries, created in the context of many agencies and without standardized lexicographic conventions.

In short, our goal is to mine the information available in human-readable glossaries made publicly available on the web. WordNet [20] is the prime example of a lexical resource that uses hierarchical relations among sets of terms to specify synonyms, hypernyms and other lexical relations. However, WordNet coverage of our glossary terminology is sparse. For example, specialized terms from our super-glossary like "radiological sabotage" or "material access center" are unlikely to appear in WordNet, which is designed to reflect ordinary English usage. Some of the genus terms linking the component glossaries bear a strong resemblance to WordNet concepts near the top of the hierarchy, e.g., "condition, material, person, process" whereas others more directly reflect the government agency domain the glossaries come from, e.g., "agency, data, document."

3. PARSING DEFINITIONS

A canonical glossary entry defines a term with respect to a superordinate category (the genus category), and provides additional information that differentiates the term from other members of the category (species). In the following definition of employee, "person" is the genus category and the phrase "works for wages in the service of" differentiates an employee from other "persons," e.g. from an employer.

(1) **Employee:**
(http://www.msha.gov/regdata/msha/56.2.htm)
Means a *person* who works for wages or salary in the service of an employer

Our current super-glossary has five entries for the term "employee" from agencies with disjoint areas of responsibility: the U.S. Department of Agriculture (USDA), the Federal Railroad Administration (FRA), the Mine Safety and Health Administration (MSHA; definition (1) above), the Office of Personnel Management (OPM), and the Securities and Exchange Commission (SEC). Of these five entries, three provide "person" or "individual" as the superordinate category. One, however, defines the term with respect to a more specific category as shown in definition (2) below ("appointed officer or employee of USDA"). In this case, one of the disjoined genus terms is in fact the same as the defined term, namely the word "employee":

(2) **Employee:**
(http://www.afm.ars.usda.gov/ppweb/34102.htm)
An appointed *officer or employee* of USDA including special Government employees (collaborators, consultants and panel members). The term excludes independent contractors.

Yet another of the definitions does not include a genus term either, as shown in (3), but this example is even more subtle:

(3) **Employee:**
(http://www.sec.gov/divisions/corpfin/forms /reg12b.htm)
The term "employee" does not include a director, trustee, or officer.

Where definition 2 mentions a subordinate category instead of a superordinate, definition 3 fails to mention a conceptual category that includes, or is included in, the term "employee. Because neither of these definitions mentions a genus category, one would have to be inferred, e.g., by creating explicit links across the multiple definitions for "employee," representing the absence of genus terms in (2) and (3), and perhaps proposing one of the genus concepts from definitions (1), (4) and (5) as the default ("person" or "individual").

(4) **Employee:**
(http://www.fra.dot.gov/counsel/regs/cfr/49/oct2000/220.htm)
means an *individual* who is engaged or compensated by a railroad or by a contractor to a railroad, who is authorized by a railroad to use its wireless communications in connection with railroad operations.

(5) **Employee:**
(http://www.opm.gov/cplmr/html/glossary)
The *term* "employee" includes an individual "employed in an agency" or "whose employment in an agency has ceased because of any unfair labor practice," but does not include supervisors and management officials or anyone who participates in a strike or members of the uniformed services or employees in the Foreign Service or aliens occupying positions outside the U.S.

Once a linked super-glossary has been created and the genus relationships established, other differentiating concepts and properties can be extracted and represented in the database. For example, as illustrated in both (2) and (3) above, three of the five "employee" entries list members of the "person … who works" genus to be "included" or "excluded" in each case. The unique concerns of the SEC are reflected in the exclusion of "director, trustee or officer" from the "employee" category.

We approach the problem of creating structure within our heterogeneous super glossary by first attempting to automatically extract the genus phrase or term from entries like (1) above. We constructed a baseline system based on shallow parsing strategies and other IE techniques [12], used human subjects to evaluate the extracte genus phrases and to define a gold standard (GS) of phrases and head words for performing a baseline evaluation. We present the results of this two-stage evaluation, and of a

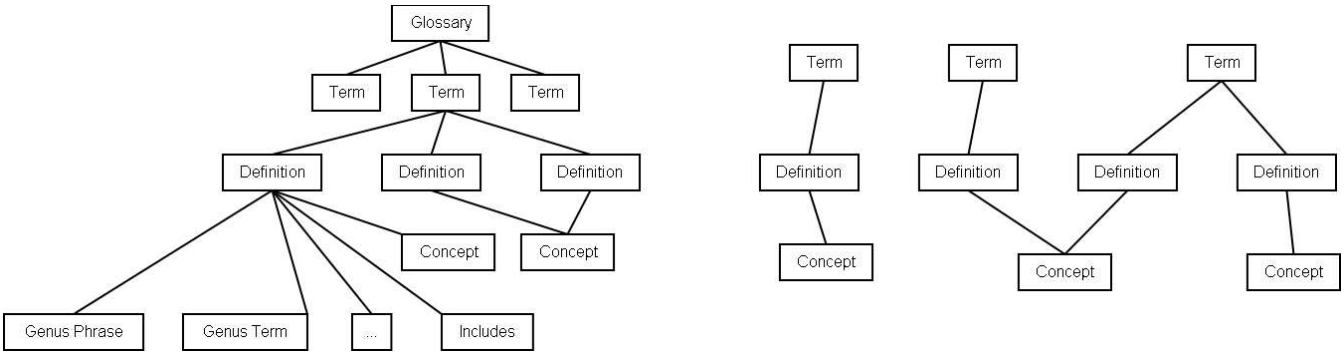

Figure 3: Two Views of the StoreGloss database

qualitative error analysis to identify the kinds of improvements in parsing that would lead to significant improvements in accuracy and coverage.

3.1 System Description

The GlossIT system shown in Figure 1 allows for automated acquisition, analysis, and linking of explicit and implicit glossaries from heterogeneous document sources on the Web. A glossary is a set of term/definition pairs, as illustrated in the examples for the term "employee" in section 2 above. An explicit glossary web page is often labeled as such. Others glossaries will not be identified as such, but conventions of layout, such as presenting term/definition pairs in a bulleted list format, can provide indirect cues to the presence of glossary material embedded within a web document containing free text, images, or other non-glossary material. GetGloss [16] locates and extracts entries from both types of glossaries. The Definder module extracts term/definition pairs from free text, a second source of "implicit" glossary data. For example, consider the following paragraph:

> (http://www.icorp.net/cardio/articles/congestv.htm)
> "The most frequent cause of the condition in older patients is atherosclerosis -- the progressive narrowing of the heart's own arteries by cholesterol plaque buildups, which starves the heart itself for oxygen and nutrients. In younger patients, it is more likely to be from a faulty heart valve or from cardiomyopathy -- damage to the heart muscle from an infection or other cause."

Definder will extract two definitions:

> **Atherosclerosis**: the progressive narrowing of the heart's own arteries by cholesterol plaque buildups, which starves the heart itself for oxygen and nutrients.

> **Cardiomyopathy**: damage to the heart muscle from an infection or other cause.

For each glossary identified by GetGloss or produced by Definder, an XML file is generated that encodes the terms and definitions from the glossary. Currently, files are stored in a directory until ParseGloss is run, although the process could be automated. ParseGloss is structured as a series of plug-in modules, which apply a series of XML transformations to the file until it is parsed into conceptual components. In early stages, shallow parsing tools are utilized (Alembic for part-of-speech (POS) tagging [1], LinkIt for noun phrase (NP) chunking [7]). Since these are run from independent modules, they can be fairly easily replaced with modules to use other tools. After that has been concluded, ParseGloss analyzes the free text definitional field to identify the genus phrase and head genus term. The NPs are considered sequentially. If an NP has an empty head, or has a head identical to the term being defined, it is skipped. The first NP not to be skipped is taken as the genus phrase. Another component of ParseGloss uses templates with cue phrases (currently 22 templates) to identify important properties of the definition. For example, the phrase "see also" is used to identify a potential cross-reference. After processing, ParseGloss produces an output XML file. A set of XSLT transformations then converts the output file into a series of formats useful for human viewing and for input into other systems.

ParseGloss loads its output into our super-glossary database, StoreGloss. The db stores the glossary terms, concept information optionally related to each term, and an arbitrary amount of additional information. It allows for two main views of data (see Figure 3): that of a forest of glossaries, and a graph of per-term information, where each term can be related to any other term. Common genus phrases may relate two otherwise distinct terms. An example might be "nonprofit", which links museum, hospital and educational institution.

3.2 Challenges

The process of identifying the genus phrase of each definition is difficult. Glossaries found on the Internet vary widely in their formatting and styles of writing. For example, consider a definition for "metallurgical coal" from the Department of Energy (DOE):

> **Metallurgical coal:**
> (http://www.fe.doe.gov/education/glossary.html)
> The type of coal which is converted to coke for use in manufacturing steel; often referred to as coking coal.

ParseGloss correctly identifies "The type of coal" as the genus phrase, and "coal" as the head genus term. Note that determining the semantics of the relationships between terms and their genus phrase is complex. While often the relationship is an IS-A relationship, in this case the definition is aimed more at defining the use of the term (that metallurgical coal is converted to coke, and used to manufacture steel), rather than identifying it as a subtype of coal.

Genus phrase identification can enable linking of definitions with similar semantics, in some cases. At the same time, contextual differences can affect the meaning of the same genus term across multiple definitions, as in the case of a relational genus term whose arguments can further restrict the meaning of the relation [5]. For example, all of the following definitions have been linked via the genus term 'examination,' but the type of examination depends on what fills the 'examinee' and 'examiner' roles.

Audit:
(www.fdic.gov)
An *examination* of the financial statements, accounting records, and other supporting evidence of an institution performed by an independent certified or licensed public accountant in accordance with generally accepted auditing standards (GAAS) and of sufficient scope to enable the independent public accountant to express an opinion on the institution's financial statements as to their presentation in accordance with generally accepted accounting principles (GAAP).

Preclosure safety analysis:
(www.nrc.gov)
means a systematic *examination* of the site; the design; and the potential hazards, initiating events and event sequences and their consequences (e.g., radiological exposures to workers and the public). The analysis identifies structures, systems, and components important to safety.

Urinalysis:
(www.pueblo.gsa.gov)
Examination of the urine for infectious agents, cells, or other substances that are signs of disease.

Industrial radiography (radiography):
(www.nrc.gov)
means an *examination* of the structure of materials by nondestructive methods,utilizing ionizing radiation to make radiographic images.

Note that the interpretation of the word "examination" must be performed in terms of the semantic domain of the database. "Examination" can have several meanings, depending on whether the domain is, for example, medical or financial.

In contrast, in many cases the genus phrase which links definitions has the same or very similar sense. For example, some of the sixty definitions with a genus of "area" are:

Ecoregions:
(www.epa.gov/waterscience/biocriteria/glossary.html)
a relatively homogeneous ecological *area* defined by similarity of climate, landform, soil, potential natural vegetation, hydrology, or other ecologically relevant variables (see also bioregions).

Island:
(http://www.access-board.gov/prowac/commrept/part3-01.htm.xml)
a defined *area* between traffic lanes for control of vehicular movements or for refuge. Within an intersection area, a median is considered to be an island.

Field:
(http://www.osha.gov/dts/sltc/methods/inorganic/id160/id160.html)
The area within the graticule circle that is superimposed on the microscope image.

4. EVALUATION

4.1 Creation of "Gold Standard"

We conducted an evaluation in order to determine the feasibility of automatic extraction of conceptual relations (specifically genus/species relations) from glossaries. We used semantic judgments from humans on an open-ended task that was related to the goal of ParseGloss. This allows us to better understand and compare possible discrepancies, differences or similarities between the commonsense way people reason about terminology and the requirements of automation. There were two steps to accomplish this. The first was to survey human respondents to determine "the most important word or phrase" that identified the term being defined. We compiled this information into a gold standard. Once the gold standard was defined, we used it to measure the accuracy of the genus phrase detection in ParseGloss. For the purposes of this paper, we have defined accuracy as the percentage correct.

4.1.1 Method

To conduct the evaluation, we first created a testing corpus containing one hundred definitions that were randomly selected from six distinct different glossaries. The glossaries spanned several domains (energy, medical, and census), and had a variety of writing styles (paragraph-length vs. full sentences vs. phrases). We then presented the testing corpus to humans in order to compile a "gold standard" of genus phrases and terms for a subset of the terms from a human annotation task.

The human annotations were collected in two phases. In the first pilot study phase using a paper questionnaire, volunteers were given two sheets of instructions, examples, and printouts of 25 definitions. They were asked to circle the "most important word or phrase" in each definition that constitutes the "heart or essence" of the definition. In addition to the main word or phrase, the volunteers were asked to identify other properties of the definition. The volunteers for this phase were both experts in searching and non-experts. The process took about one hour. Volunteers were not paid.

The second phase was initiated at the request of the volunteers of the first phase. They observed that a lot of time was taken by simply writing out the information already present on the paper. A second phase was then begun, where subjects were given access to a browser-based interface[1] where they could provide the same information by means of cutting and pasting instead of

[1]http://www.cs.columbia.edu/digigov/LEXING/eval/instructions.html

rewriting. We lowered the number of definitions per session to five, in the expectation that a fifteen minute evaluation would attract more participants. The volunteers for this phase were from the same groups as those for the first phase. Volunteers were not paid.

There were a total of 7 respondents for the paper-based evaluations and 26 respondents for the web-based evaluation. Each term was judged by at least 3 people, up to 5 in total.

In addition to the second phase, we also solicited four volunteers to participate in a "think-aloud" style experiment. They performed the same task as those in the second phase, but while they were reading the definitions and making their choices, they were asked to speak out loud and explain what they were thinking and why they made their choices. Their thoughts were recorded on a tape recorder. The four volunteers consisted of three area specialists and one layman. Results of this task will be reported in future work.

4.1.2 Results

We have analyzed the results of the evaluation. Specifically, we have analyzed the genus phrases chosen by the human subjects, and the head words of the phrases that they have chosen.

Given that it is difficult to get agreement even from carefully trained linguists on syntactic constituency, (note the detailed rules needed for the Penn Treebank annotations by graduate students in linguistics [19]), we hypothesized that we might not get high levels of agreement among subjects on the exact strings, but we did expect good agreement on the region of text selected. For both exact agreement and region selection, results were as hypothesized. In addition, some words do not convey much extra meaning. For example, when discussing measurements, the difference in meaning between a "unit" and a "standard unit" is small and subtle; we hypothesized that examples like this would have the effect of increasing the likelihood of subjects' disagreement on whether the word "standard" would be part of the response. However, we would not expect people to give completely disjoint responses. This is consistent with the results, and the relatively high agreement for Type V agreement, as described below.

One challenge is that the subject who took the evaluation on paper often would rephrase the definition, taking words from different places in the definition. For example, consider the definition:

> **Drilling mud:**
> (http://www.fe.doe.gov/education/glossary.html)
> A special mixture of clay, water, or refined oil, and chemical additives pumped downhole through the drill pipe and drill bit. The mud cools the rapidly rotating bit; lubricates the drill pipe as it turns in the well bore; carries rock cuttings to the surface; serves as a plaster to prevent the wall of the borehole from crumbling or collapsing; and provides the weight or hydrostatic head to prevent extraneous fluids from entering the well bore and to control downhole pressures that may be encountered.

One response from the paper task was a "cooling mixture", a phrase which does not occur in the original definition, but is the result of sophisticated human synthesis. At this time, such synthesis is beyond the scope of our techniques. Since we want to automatically identify a phrase from the definition, we excluded such responses when they were not applicable, but included them whenever possible (e.g. they were always counted for Type V, onset, and ending).

In analyzing the responses, we defined several types of agreement as shown in Table 1. Type I is the most strict, corresponding to agreement on the exact string of words selected from definitions by respondents. From Type I to Type V, portions of the string are successively omitted from consideration, thus Type II excludes relative clauses. Type V is the least inclusive, consisting only of the head of the phrase. The last three types indicate whether subjects agreed on the first word, last word, or a subset of words in the string.

Table 1: Types of Agreement

Type	***What is included***
Type I	The full string is matched, with the exception of determiners, punctuation, and terms with almost no semantic information. Percentage agreement.
Type II	The string is matched, excluding relative clauses, and the items from Type I. Percentage agreement.
Type III	The string is matched, excluding prepositional phrases, and the items from Type II. Percentage agreement.
Type IV	The string is matched, excluding premodifiers and the items listed in Type III. Percentage agreement.
Type V	Only the headword is matched. Percentage agreement.
Onset	Matched if the different subject chose the same place within the definition to begin the phrase. Percentage agreement.
Ending	Matched if the different subjects chose the same place to end the phrase. Percentage agreement.
Inclusion	Matched if the gold standard term was entirely included in the response. Percentage agreement.

In general, we discounted words that had less semantic information. For example, "barrel" is defined as "The standard unit of measure of liquids in the oil industry." When defining a unit, there is little difference between a "unit" and a "standard unit". Therefore, we counted both responses the same. This affected four terms, listed in Table 2.

The results are summarized in Table 3. Note that agreement among subjects for Type I indicates that a majority of the subjects agreed on the exact string selected to be "the most important" part of the definition most of the time. More importantly, the high agreement on Type V and inclusion indicate that subjects nearly always found the same region of text.

Table 2: Definitions with Merged Responses

Term	*Def*	*Responses*
Barrel	The standard unit of measure of liquids in the oil industry; it contains 42 U.S. standard gallons.	1) standard unit of measure 2) unit of measure
BLOWOUT	An uncontrolled flow of gas, oil, or other fluids from a well into the air. A well may blow out when pressure deep in the reservoir exceeds the weight of the column of drilling fluid inside the well hole.	1) uncontrolled flow 2) uncontrolled flow of gas, oil, or other fluids
btu per cubic foot	The total heating value, expressed in Btu, produced by the combustion, at constant pressure, of the amount of the gas that would occupy a volume of 1 cubic foot at a temperature of 60 degrees F if saturated with water vapor and under a pressure equivalent to that of 30 inches of mercury at 32 degrees F and under standard gravitational force (980.665 cm. per sec. squared) with air of the same temperature and pressure as the gas, when the products of combustion are cooled to the initial temperature of gas and air when the water formed by combustion is condensed to the liquid state. (Sometimes called gross heating value or total heating value.) OPI EI-40 Sources FERC-2	1) Heating value 2) total heating value
collection block	The smallest area that the U.S. Census Bureau used to collect information for the decennial census. A collection block may be split by the boundary of anylegal or statistical entity later recognized by the Census Bureau for census data presentation.Thus, if a collection block is split by one or more legal and/or statistical boundaries, each portion will be a separate tabulation block; if a collection block is not split, the same area maybe a tabulation block. See block number, census block, tabulation block.	1) smallest area that the U.S. Census Bureau used to collect information for the decennial census 2) Smallest area used to collect info. For the decennial census

Table 3: Overall Agreement

Type	*Agreement*
Type I	0.54
Type II	0.61
Type III	0.74
Type IV	0.83
Type V	0.87
Onset	0.78
Ending	0.63
Inclusion	0.81

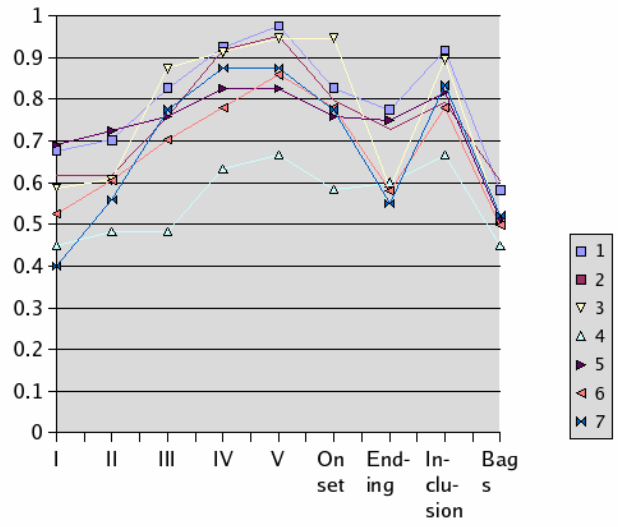

Figure 4: Human Agreement by Glossary

Table 4: Glossary sources

1	http://www.fe.doe.gov/education/glossary.html
2	http://www1.cs.columbia.edu/~smara/DEFINDER/
3	http://www.epa.gov/OCEPAterms/aterms.html
4	http://www.cs.columbia.edu/nlp/flkb/eia_small/gloss.html
5	http://www.epa.gov/OCEPAterms/aterms.html (second group)
6	http://www.census.gov/geo/www/tiger/glossary.html
7	http://www.cdc.gov/nccdphp/drh/epi_gloss.htm

Figure 4 summarizes the results by glossary. The glossaries are listed in Table 4. Note that the results vary greatly by glossary number, but the overall trend of increasing agreement from Type I through Type V persists.

4.1.3 Discussion

On an open-ended task of locating "important" phrases within definitions, human subjects perform as expected: they exhibit variation on the exact string selected, even discounting semantically empty words like "unit," but show very high agreement on the region of text identified, as reflected in both the increasing agreement in Table 2 from Types I through V, and on the high agreement for "inclusion" (81%). Subjects agree more on onsets of phrases (78%) than endings (63%), which reflects the right-branching syntactic structure of English, and the cognitive salience of word, phrase and sentence onsets in such languages.

From the human responses, we created a gold standard based on "consensus": by looking for simple majority agreement from Type I. If none existed, we selected the majority agreement from the Type: Inclusion. If that did not produce a majority, we chose the shortest item. Examples of this were the terms "Associated gas" and "Absorption", as listed in Table 5. In three cases of the 100, there was no agreement and we elected to drop those items from the gold standard.

Table 5: Responses Using the Inclusion Rule for GS Selection

Term	*Def*	*Responses*
Associated gas	Gas combined with oil. Known also as gas cap gas and solution gas, it provides the force (also called the drive mechanism) needed to force oil to the surface of a well. Associated gas is normally present in an oil reservoir in the early stages of production.	1) Gas 2) Gas combined with oil
Absorption	The uptake of water , other fluids, or dissolved chemicals by a cell or an organism (as tree roots absorb dissolved nutrients in soil.)	1) uptake of water 2) uptake of water, other fluids 3) uptake of water, other fluids, or dissolved chemicals

From the gold standard of 97 genus phrases, we also identify the head words and evaluate the performance of ParseGloss in extracting both components from free text.

4.2 Baseline System Performance

4.2.1 Method

4.2.2 Results

Having compiled the gold standard, we then ran ParseGloss to compare the results. The genus phrases gave an accuracy of 59%, which is higher than the human agreement (type I); the head terms gave an accuracy of 64%, reflecting our use of shallow parsing techniques. This focus was due to the focus on broad coverage. In our other components (Definder) we use a combination of deeper linguistic analysis and shallow techniques and in future work we may extend this to the ParseGloss module. This suggests a need for full parsing in order to extract reliable semantic information (cf. similar point in [9]).

Having a full parse of the definitions would allow us to write more complex rules which allow for certain cases not supported by the current shallow parsing. For example, the EPA defines:

Acclimatization:
(http://www.epa.gov/OCEPAterms/aterms.html)
The physiological and behavioral adjustments of an organism to changes in its environment.

The sentence is parsed as follows:

<s><lex pos=DT>The</lex> <lex pos=NN>physiological</lex> <lex pos=CC>and</lex> <lex pos=JJ>behavioral</lex> <lex pos=NNS>adjustments</lex> <lex pos=IN>of</lex> <lex pos=DT>an</lex> <lex pos=NN>organism</lex> <lex pos=TO>to</lex> <lex pos=NNS>changes</lex> <lex pos=IN>in</lex> <lex pos="PRP$">its</lex> <lex pos=NN>environment</lex><lex pos=".">.</lex></s>

Which leads to the following Noun-phrase (NP) output:

<NP>The physiological</NP> and <NP>behavioral adjustments</NP> of <NP>an organism</NP> to <NP>changes</NP> in <NP>its</NP> <NP>environment</NP>.

Since a major component of the parsing algorithm is to identify the first NP, the genus phrase identified using this approach is "The physiological". One reason for the mistake is that "physiological" is mis-tagged as a noun instead of an adjective. The gold standard for this term is "adjustments", which is much more sensible. The Collins parser [4] outputs the following parse tree:

<TOP> <NP> <NPB> <lex pos="DT">The</lex> <lex pos="NN">physiological</lex> <lex pos="CC">and</lex> <lex pos="JJ">behavioral</lex> <lex pos="NNS">adjustments</lex> </NPB> <PP> <lex pos="IN">of</lex> <NP-A> <NPB> <lex pos="DT">an</lex> <lex pos="NN">organism</lex> </NPB> <PP> <lex pos="TO">to</lex> <NP-A> <NPB> <lex pos="NNS">changes</lex> </NPB> <PP> <lex pos="IN">in</lex> <NP-A> <NPB> <lex pos="PRP">its</lex> <lex pos="NN">environment</lex> <lex pos="PUNC.">.</lex> </NPB> </NP-A> </PP> </NP-A> </PP> </NP-A> </PP> </NP> </TOP>

The POS stage parses the beginning of the sentence identically. However, the emphasis on maximal noun phrases instead of simplex noun phrases led to a longer initial NP, with a head of "adjustments", instead of "physiological". More importantly, this breakdown allows a rule to identify the first few words as modifying the head noun. We can then apply a more general rule to ignore premodifiers.

A summary of rules to parse glossary entries using the Collins parser [4] is listed in Table 6

Table 6: Proposed Rules Using Collins Parser

Rule/Pattern	*What to do*	*Description*
Always	Select text from the beginning	Where to start
Always	Stop selecting at SBAR, VP, PP, etc	Where to end
Always	Remove initial DT	Initial Determiners
1st NP contains only DT	Skip to next PP	Empty head
Head of 1st NP is term being defined	Skip to next PP	Empty head
Head of 1st NP is one of (Type, ...)	Skip to next PP	Empty head
Head of 1st NP is one of (unit, measure...)	Include the next PP	Semi-empty head
Usually (see next)	Remove JJ, etc before 1st NP	Premodifiers
JJs before head noun one of (uncontrolled, special, dry, ...)	Keep JJs	Important premodifiers

4.2.3 Discussion

There were several factors that led to the 59% accuracy of ParseGloss. The most significant is definitions that confuse the version of the tagger (Alembic 2.8) that we are using. Commas are interpreted as delimiting the ends of phrases, and hence cannot identify a comma-separated list of items as a single

constituent; this type of construction, however, is common in glossaries: cf. our definition of 3. of employee from section 2 above: *The term "employee" does not include a director, trustee, or officer.* Also, there are some definitions that the current combination of tools was unable to parse. ParseGloss was then unable to output a genus phrase, and this lowered the accuracy. In order to combat both problems, we propose testing tools with another combination, such as the Collins parser [4] or Charniak parser. In addition, using the full information provided by other parsers instead of the current shallow parsing of NPs would allow us to built a set of rules to increase the accuracy.

For example, parsing would begin by determining the type of sentence. For most sentences, parsing would start with the first NP (ignoring NPs that enclose the entire sentence.) Initial determiners would be stripped off. ParseGloss would extract all elements until it encountered certain elements (VP, SBAR, most PPs). Some sentence types, such as those of the form "VP NP" would be handled differently – in those cases, VP is usually intended, not the NP.

Very long encyclopedic-style definitions were more difficult to parse. We therefore propose to treat the very long entries differently. In addition, there is no uniformity regarding when to include trailing prepositional phrases (PP). For example, (GS italicized):

Blowout:
(http://www.fe.doe.gov/education/glossary.html)
An *uncontrolled flow* of gas, oil, or other fluids from a well into the air. A well may blow out when pressure deep in the reservoir exceeds the weight of the column of drilling fluid inside the well hole.

Barrel:
(http://www.fe.doe.gov/education/glossary.html)
The standard *unit of measure* of liquids in the oil industry; it contains 42 U.S. standard gallons.

In the first case, no trailing PPs are desired, but the second, the first PP is desired, but not the following ones. Sometimes, a PP beginning with "of" is desired, but often it is not. Often, though, an indicator of this is the head noun of the PP, such as the empty head noun. We suggest that this list may be obtained by querying the database to find the most common nouns preceding PPs. For example, in this set, we discovered that the word "unit" is a good predictor. In addition, there are cases where the head of the NP might imply that the NP is completely empty, and processing should begin with the next phrase. In this set, the word "type" is a good predictor. In addition, any time the head of the NP is a determiner (e.g. "this"), or the same as the term being defined, processing should continue with the next phrase.

5. FUTURE WORK

Future work includes three directions. First, in order to further explore the evaluation methods presented, we would like to test a larger number of definitions with more subjects. This would enable us to divide definitions into different categories such as types (encyclopedic, phrasal, etc), domains, and language level. A better understanding of the impact of definition type on the ability of humans to perform the task of identification of semantic components of definitions would permit us to establish a clear upper and lower bound for the ParseGloss algorithm.

The second direction is to change the toolset used by ParseGloss to explore the impact of different analyses on performance. In our current implementation we are using the Alembic Tagger [1] (http://www.mitre.org/technology/alembic-workbench/) and Link-IT noun phrase chunker [7]. However, we have observed errors such as the following (from the DOE's Fossil Fuel Education Glossary).

Electrostatic Precipitator:
(http://www.fe.doe.gov/education/glossary.html)
An electrically charged device for removing fine particles (fly ash) from combustion gases prior to the release from a power plant's stack. The device passes combustion gases through positively and negatively charged plates that attract the tiny particles using static electricity.

In this example, the head term "electrostatic precipitator" is carefully defined, but the tagger mislabels ``electrically" as a noun, so ``An electrically" (shown in italics) is thus tagged as a noun phrase. This is an error due to the presence of ``charged" which is labeled as a verb, and thus forces whatever comes before it to be a noun. When Link-IT processes the definition, the genus phrase is chosen to be ``electrically", rather than ``an electrically charged device". This is clearly incorrect. These errors will be addressed as part of the evaluation and post-processing modules will be built to correct them.

Finally, since our ultimate goal is to use the output of our analyses as part of an ontology for knowledge representation, we will explore ways to incorporate our output into existing ontologies, for example the Omega ontology [11]. We will seek partners for whom populating an ontology for semantic access will be helped by having the data provided by ParseGloss.

6. ACKNOWLEDGMENTS

This research was supported by NSF Funding in the Digital Government program.

7. REFERENCES

[1] Aberdeen, John; Burger, John; Day, David; Hirschman, Lynette; Robinson, /Patricia; Vilain, Marc. 1995. "MITRE: Description of the Alembic System as used for MUC-6". In Proceedings of the Sixth Message Understanding Conference (MUC-6), 1995.

[2] Berners-Lee, T., Hendler, J., Lassila, O. 2001. "The Semantic Web". Scientific American, May 2001.

[3] Byrd, Roy J., Nicoletta Calzolari, Martin Chodorow, Judith L. Klavans, Mary S. Neff and Omneya A. Rizk (1987). "Tools and Methods for Computational Lexicology". *Computational Linguistics* 13(3-4): 219-240.

[4] Collins, Michael. 1996. "A New Statistical Parser Based on Bigram Lexical Dependencies". Proceedings of the 35th Annual meeting for Computational Linguistics, Santa Cruz, CA

[5] Dahl, D.; Palmer, M.; Passonneau, R. J. Nominalizations in PUNDIT. Proceedings of the 25th Association for Computational Linguistics. Stanford, CA. 1987.

[6] Ding, Ying; Foo, Schubert. 2002. Ontology Research and Development: Part 1 – A Review of Ontology Generation. Journal of Infromation Science 28(2).

[7] Evans, David K., Judith L. Klavans and Nina Wacholder (2000). "Document Processing with LinkIT". RIAO 2000, Recherche d'Informations Assistee par Ordinateur. Paris, France, pp. 1336-1345.

[8] Fellbaum,, Cristiane, ed. 1998. WordNet: An Electronic Lexical Database. MIT Press.

[9] Gildea, Daniel and Martha Palmer. 2002. The necessity of syntactic parsing for Predicate Argument Recognition. Proceedings of the 40th Annual Conference of the Association for Computational Linguistics, Philadelphia, PA.

[10] Guthrie, L., Guthrie, J., Wilks, Y., Cowie, J., Farwell, D., Slator, B. and Bruce, R. (1992). A research program on machine-tractable dictionaries and their application to text analysis. Literary and Linguistic Computing, vol. 8, no. 4 (special issue, eds. Ostler and Zampolli)

[11] Hovy, E.H; Fleischman, M.; Philpot, A. "The Omega Ontology". In progress.

[12] Jacobs, Paul S., ed. 1992. Text-Based Intelligent Sytems : Current Research and Practice in Information Extraction and Retrieval. Hillsdale, NJ: L. Erlbaum Associates.

[13] Klavans, Judith L. (1994). "Visions of the Digital Library: Views on Computational Linguistics and Semantic Nets in Information Retrieval". Festschrift for Donald E. Walker. Antonio Zampolli, Nicoletta Calzolari and Martha Palmer, editors. Kluwer, New York.

[14] Klavans, Judith L., Martin S. Chodorow and Nina Wacholder (1990). "From Dictionary to Knowledge Base via Taxonomy". Proceedings of the Sixth Conference of the University of Waterloo Centre for the New Oxford English Dictionary and Text Research: Electronic Text Research. University of Waterloo: Waterloo, Canada.

[15] Klavans, Judith L. and Smaranda Muresan (2001a) "Evaluation of DEFINDER: A System to Mine Definitions from Consumer-oriented Medical Text". Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL). Roanoke, Virginia, pp. 201-203.

[16] Klavans, Judith L., Davis , Peter T. and Popper, Samuel (2002). "Building Large Ontologies using Web Crawling and Glossary Analysis Techniques". Proceedings of the National Conference for Digital Government Research. Los Angeles, California.

[17] Lin, S. H., Shih, C. S. , Chen, M. C., Ho, J. M., Kao, M. T., and Huang, Y. M., "Extracting Classification Knowledge of Internet Documents: A semantics Approach", ACM SIGIR'98, 1998, pp. 241-249.

[18] Maedche, A., Staab, S.. Ontology learning for the Semantic Web. IEEE Intelligent Systems, 16(2), 2001.

[19] Marcus, Mitchell P.; Santorini, Beatrice; Marcinkiewicz, Mary Ann. 1993. Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics 19.2.

[20] Miller, George A.; Fellbaum, Christiane; Kegl, Judy; Miller, Katherine J. 1988. WordNet: An electronic lexical reference system based on theories of lexical memory. In Revue quebecoise de linguistique 17(2), 1988, pp. 181-213.

[21] Muresan, Smaranda, and Judith L. Klavans (2002). "A Method for Automatically Building and Evaluating Dictionary Resources". Proceedings of the Language Resources and Evaluation Conference (LREC 2002). Las Palmas, Spain.

[22] Procter, Paul (1978) Longman Dictionary of contemporary English, Longman Group Limited, Harlow and London, England

[23] Sinclair , John M., ed. Collins COBUILD English Language Dictionary. Collins, London, 1987. Web site: http://titania.cobuild.collins.co.uk/.