An NLP Approach for Improving Access to Statistical Information for the Masses

Elizabeth D. Liddy & Jennifer H. Liddy
Center for Natural Language Processing
School of Information Studies, Syracuse University

Abstract

Naïve users need to access statistical information, but frequently do not have the sophisticated levels of understanding required in order to translate their information needs into the structure and vocabulary of sites which currently provide access to statistical information. However, these users can articulate quite straightforwardly in their own terms what they are looking for. One approach to satisfying the masses of citizens with needs for statistical information is to automatically map their natural language expressions of their information needs into the metadata structure and terminology that defines and describes the content of statistical tables. To accomplish this goal, we undertook an analysis of 1,000 user email queries seeking statistical information. Our goal was to better understand the dimensions of interest in naïve users' typical statistical queries, as well as the linguistic regularities that can be captured in a statistical-query sublanguage grammar. We developed an ontology of query dimensions using this data-up analysis of the queries and extended the ontology where necessary with values from actual tables. We proceeded to develop an NLP statistical-query sublanguage grammar which enables the system to semantically parse users' queries and produce a template-based internal query representation which can then be mapped to the tables' metadata, in order to retrieve relevant tables which are displayed to users with the relevant cell's value highlighted.

Introduction

With the ever-increasing availability of government-provided statistical information on the Web, there exists a very substantive need and responsibility to provide useful, valid information to users who have an unsophisticated level of statistical literacy, but who do have a need to access and utilize statistical information provided by government agencies. The Citizens' Access to Statistical Data: A Study of Tabular Data (http://istweb.syr.edu/~tables/), funded by the NSF Digital Government Initiative, is a wide-ranging project that has been: 1) investigating *why* and *how* people seek statistical information, and; 2) developing and testing prototype tools that aid in finding, displaying, and utilizing statistical information found in tables. Our goal has been to gain a sufficient understanding of this population which, in turn, will enable us to develop high-quality information access technologies that can be utilized by government agencies that bear the responsibility for providing statistical information to the general public.

In this paper, we focus on one aspect of the research we have done within this project which focused on better understanding and assisting naïve users in finding their requested information by getting them to the right table or set of tables. Within our project there are two approaches for achieving this goal. One is to use graphical browsing tools - an approach that has been investigated by Gary Marchionini (Marchionini, 1999) and Ben Schneiderman (Tanin & Shneiderman, 2001). The other is to empower users to ask their questions quite naturally, the same way they do when asking a reference librarian or when submitting email queries to a virtual reference service, but with the added benefit of dynamic interaction with the answer-providing statistical tables. This querying approach, which will be the focus of this paper, uses Natural Language Processing (NLP) to interpret and represent a user's need and to match this representation against the metadata representation of tables' contents in order to find the requested information.

In this paper, we will: motivate the study; present the methodology used to discover the frequently occurring dimensions in users' queries; describe the process of developing a statistical-query sublanguage grammar; present examples of how the grammar represents users' queries, and; show the mapping of query dimensions to table metadata. We will conclude with our views of how these results can be utilized in providing citizens with access to the statistical information to which they are entitled, and some possible future work to accomplish this goal.

Motivating Problem & Proposed Solution

The team participating in our NSF-funded Digital Government project has rooted its work in a search for better understanding and support of citizens and their needs for access to statistical data (Hert, 1999; Hert et al, 2000; Marchionini et al, 2001). We found that while much attention in the past has been paid to the needs, preferences, and practices of individuals whose occupations require almost daily accessing of statistical information, we wanted to focus instead on the masses – that is, the remainder of citizens whose daily work is not involved with governmental statistical information, but who may once in a while, have a requirement for statistical information. These users are nowhere near as familiar with the processes by which statistical information is collected, organized, and made accessible. But they are quite aware of the specifics of their own needs. So rather than focusing on the *data*, which is well-understood by professional users of statistical information, but which is a virtual unknown to much of the general public, the research we are herein reporting on, focused on understanding the information needs of *everyday users* – a group to which most US citizens belong.

It is known from years of research in the fields of librarianship and reference, that users can either be successfully guided or seriously distracted from their real search by having to interact with pre-conceived choices, steps, and options, that they are forced to make in an application in which the system is in charge. Many of these steps force users into options that focus on parameters that do not pertain to dimensions of importance in the user's query. That is, the options from which the user is asked to select are remote from the aspects of the topic about which they are inquiring. While human intermediaries have the ability to interact with users in such a way as to facilitate the user's basic need becoming known and presented in the most accurate and descriptive terms, computer interfaces can lead users through a series of choices which may be only minimally related to the real need of the user, and which, in fact, end up failing to respond to the real intent of the user's query.

The most powerful solution to this problem is to allow users to express quite straightforwardly in natural language sentences what they are looking for, and to provide even more context if so moved. While it is true that many search engines today do allow users to enter their own description of their need, both the size of the query box and the examples used by commercial search engines, mitigate against users' straightforwardly expressing their full information need. It is also the case that the great majority of commercial search engines do not know how to deal with the implicit and explicit information that is encoded in natural language queries, and so even if they encouraged users to enter fully expressive natural language queries, the search engines would simply reduce them to space delimited tokens. Our approach is to utilize Natural Language Processing (NLP) to represent users' queries and correctly map them to the metadata used to represent the content of statistical tables

Natural Language Processing

NLP is a range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of particular tasks or applications (Liddy, 1998). The possible levels of linguistic analysis are:

- Morphological - componential analysis of words, including prefixes, suffixes and roots
- Lexical - word level analysis including lexical meaning and part of speech assignment
- Syntactic - analysis of words in a sentence in order to uncover the grammatical structure of the sentence
- Semantic - determining the possible meanings of a sentence, including disambiguation of words in context
- Discourse - interpreting structure and meaning conveyed by texts larger than a sentence
- Pragmatic - understanding the purposeful use of language in situations, particularly those aspects of language which require world knowledge

There are two central tasks for NLP in providing users with the statistical information they seek. First is the translation of potentially ambiguous natural language queries. Second is the representation of answer-providing sources in an unambiguous internal representation on which matching and retrieval can take place. In fact, the ideal Information Retrieval (IR) system is one in which users express their information needs naturally and with all requisite detail - exactly as they would state them to a research librarian. The system would then "understand" the underlying meaning of the query in all its complexity and subtlety. Furthermore, the ideal IR system would represent the contents of documents - no matter the nature of the document - at all the same levels of understanding, thereby permitting full-fledged conceptual matching of queries and documents.

Sublanguage Grammar

Within the field of NLP, our research made substantive use of the theoretical and empirical methodology of Discourse Linguistics, the specialization concerned with understanding how different communication types convey meaning. The discourse level model of a 'communication-type' consists of a particular set of components of information and relations among these components. Discourse characteristics are used by humans, and can be simulated by an NLP system, to interpret levels of meaning beyond the simple surface level

Within the discipline of Discourse Linguistics, we have developed a sublanguage grammar that recognizes the distinct structure and semantic content of queries seeking statistical information. Research in Sublanguage Theory (Sager, 1981; Sager et al, 1987; Liddy, 1991; Liddy et al, 1993) has shown that communication types that are used for a common purpose exhibit characteristic lexical, syntactic, semantic, discourse, and pragmatic features. A sublanguage grammar reflects the *information structure* of the domain, while the semantic classes of words used and the semantic relations between these classes reflect the *knowledge structure* of the domain. The process of developing a sublanguage grammar for a particular genre is a data-centered approach to knowledge representation and results in a well-grounded domain model which provides guidance in learning the particularized linguistic rules for both understanding the meaning of text expressed in this sublanguage, and then developing technology to simulate this understanding (Liddy et al, 1993). Text types which have been analyzed and grammars developed include abstracts, news articles, arguments, instructions, manuals, dialogue, instructions, and queries (Liddy et al, 1993).

The work we are herein reporting has focused on the development of a statistical-query engine which takes as input any natural language inquiry regarding statistical information and by application of the statistical query sublanguage grammar produces a query structure which reflects the appropriate logical combination of the semantic requirements of the question. The basis of the 'query constructor' is a sublanguage grammar that is, in turn, a generalization over the regularities exhibited in the natural language expressions of sample queries analyzed in this study. The query constructor utilizes pattern-action rules to convert a query into a first order logic representation, reflecting the appropriate semantic expansion and logical organization of the content of the query. This representation is then available to the search engine for mapping into the metadata representation of statistical table elements.

Methodology

Our work, which reflects a typical empirical discourse linguistic approach, consisted of the following steps:

1. Review the sample of queries.
2. Separate out those that are not requests for statistical information.
3. Analyze remaining queries to detect their common underlying dimensions.
4. Develop an ontology of the dimensions of users' queries.
5. Fill in ontology as needed from the tables themselves.
6. Analyze the queries to detect the frequently occurring syntactic structures of how dimensions are ordered and lexicalized.

7. Write a grammar that captures these orders, choices, and variations.
8. Map statistical query dimensions into the typical labels identifying tables, columns, rows, sub-rows, and cells.
9. Test whether the grammar accurately covers a new test set of queries.

The methodology we selected enabled us to better understand: 1) what users are asking about; 2) how they ask their queries, so that we could capture it in a query sublanguage grammar, and finally; 3) how NL queries can be used for retrieval by mapping users' query dimensions onto tables' metadata elements.

We operationalized the first research question by applying human content analysis techniques to 1,000 actual user email queries from the logs of government agencies that provide statistical information on the web. The queries were manually analyzed in order to better understand the dimensions of interest in statistical queries, as well as the linguistic regularities that need to be captured in a statistical-query sublanguage grammar. The goal at this step was to determine the typical dimensions of users' queries, and to enable us to go on to the development of an ontology of query dimensions.

However, before commencing with this analysis, we needed to weed out those queries that were not really requests for statistical information. These fell into 6 main categories:

1. Requests for some action to be taken by the agency
2. Questions about the manipulation of data
3. Questions about the availability of certain data
4. Questions about data collection
5. Vague requests
6. Questions about a specific piece of data

Ontology

Based on the remaining inquiries, we developed an ontology of query dimensions using a data-up analytic approach. The top level dimensions reflect an abstraction from the specifics mentioned in the queries – and capture the various aspects relating to a statistic that are much as one might expect – they reflect aspects of the common WHO, WHERE, WHEN, WHAT aspects of journalistic reporting. They are devoid of the WHY and HOW aspects, which are included in the six types of queries which we did not analyze in this particular research project. Figure 1 shows a synopsis of the top four levels of the ontology's dimensions.

We extended the ontology at some points with values from actual tables to get a better representation of the range of possible values of a particular dimension (e.g. occupations). The full ontology can be thought of as a rich description of the aspects of data for which statistical information is sought by users. It reflects the aspects, if not all the particulars, of what users inquire about and reflects the conceptual organization gleaned from queries themselves rather than the structure imposed by those who do the data collection, organization, and aggregation.

- WHO
  - Ethnicity
    - African
  - Age
    - Years of age
      - Under the age of 15
  - Gender
    - Sex
  - Religion
    - Protestantism
      - Baptist
  - Households
    - Family households
    - Homeless
- WHERE
  - Location
    - Region
      - New England
    - State
      - New York
    - County
      - Onondaga
    - City
      - Syracuse
    - Census Tract
- WHEN
  - Time
    - 1990's
    - present
    - January
- WHAT
  - Income
  - Education
    - Status
      - Level of Education
  - Economic Indicators
    - Consumer Expenditures
    - Unemployment Rate
  - Employment
    - Full-time
  - Occupation
    - While Collar
      - Financial Managers

Fig. 1: Ontology of statistical query dimensions

Statistical Query Grammar

The next step was to get some sense of how the ontology's dimensions are reflected in the substantive content of users' queries. This revealed the 'semantic grammar' of queries, which would then, in turn, be more fully detailed with a lexico-syntactic grammar that reflects how the semantic dimensions are realized by lexical, part of speech, and ordering choices. The semantic grammar revealed that there are some combinations of dimensions that are more typical than others. For example, the WHO + WHAT + WHEN pattern accounted for 20.7% of the queries, followed by the WHAT + WHERE + WHEN pattern which accounted for 12.8%. The full details of this stage of the sublanguage grammar development as well as the lexico-syntactic patterns which constitute the statistical-query grammar which will be reported in much greater detail in a computational linguistic paper (Liddy & Liddy, forthcoming). Some sample email queries for the WHO + WHAT + WHEN pattern are:

> *"I am trying to find the percentage of women in the workforce from the years 1900 to 1998."*
>
> *"I want to know how many people worked for small businesses last year."*
>
> *"I was wondering if you might be able to send me the percentage of the working-age male population in the workforce from 1950 to present."*

> *"Could you send me stats on African Americans in white-collar jobs from 1960-98?"*
>
> *"What was the average amount of time women spent on housework per week in 1900; 1950; 1995?"*
>
> *"I seem to be having trouble finding statistics on the percentage of high school students who belong to the work force during the school year."*

Query Processing

To exemplify what occurs when the NLP statistical-query sublanguage grammar processes a user's query and produces an internal query representation based on the dimensions of the ontology, consider the following example query and the stages of processing it undergoes:

> *"In 1996, how many years was a 50 year old woman from the US expected to live?"*

After part-of-speech tagging:

> *In*|IN *1996*|CD *,*|, *how*|WRB *many*|JJ *years*|NNS *was*|VBD *a*|DT *50*|CD *year*|NN *old*|JJ *woman*|NN *from*|IN *the*|DT <CTRY> *US*|NP </CTRY> *expected*|VBD *to*|TO live|VB ?|?

After the query grammar has been applied:

> *In* <WHEN> *1996* </WHEN>, <HOW MUCH> *how many years* </HOW MUCH> *was a* <WHO> *50 year old woman* </WHO> *from the* <WHERE> <CTRY> *U.S.* </CTRY </WHERE> <WHAT> *expected to live* </WHAT> ?

The new element seen in this output from the query analyzer, is 'HOW MUCH'. This represents the quantification being sought, and is the place holder for the statistical answer. It is present in various phrasings in queries, but mainly as HOW MUCH and HOW MANY.

Another example may further clarify the labeling of the output of the query – this one omits the part-of-speech stage of the system's tagging.

> *"How many black women living in New York City in 1999 were unemployed?"*
>
> <HOW MANY> *How many* </HOW MANY> <WHO> *black women* </WHO> <WHERE> *living in New York City* </WHERE> <WHEN> *in 1999* </WHEN> <WHAT> *were unemployed* </WHAT>?

Linking to Table Elements

Having developed the basic structure of the query grammar, we needed to determine how the query dimensions would map into the structure and description of the statistical tables – the sources of answers. This is a difficult task to accomplish since there is such variety in tables both in and across agencies. While other groups (e.g. Hert, 2001) are approaching the problem by starting at the metadata end of the spectrum, and we have begun at the opposite end, that of the dimensions of users' queries, a mapping tool is needed to bridge the two approaches. Our explorations have focused on utilizing the ontology of query dimensions to perform this function. We have progressed in our work to the point of mapping query dimensions into table labels – either column, row, sub-row, cell, or table – as a possible representation of the metadata values of each. Obviously, each of these elements is defined in most instances by more detail accompanying the table, still it appears that the task of getting users to the right table and cell may be accomplished in many cases by mapping into labels. This can be shown using the two sample queries presented above by mapping the content of the generic query dimensions of WHO, WHAT, WHERE, and WHEN to the more specific levels in the ontology, which are themselves the vocabulary used as table, column, row, sub-row, and cell labels. The X indicates the value of the cell that is being sought.

> *In* <WHEN> *1996* </WHEN>, <HOW MANY> *how many years* </HOW MANY> *was a* <WHO> *50 year old woman* </WHO> *from the* <WHERE> <CTRY> *U.S.* </CTRY </WHERE> <WHAT> *expected to live* </WHAT> ?
>
> <WHEN> *01/1996-12/1996* </WHEN>, <HOW MUCH> *X* </HOW MUCH> <WHO> *female, 50-years-old* </WHO> <WHERE> *United States* </WHERE> <WHAT> *Life Expectancy*</WHAT> ?

and for query 2:

> <HOW MANY> *How many* </HOW MANY> <WHO> *black women* </WHO> <WHERE> *living in New York City* </WHERE> <WHEN> *in 1999* </WHEN> <WHAT> *were unemployed* </WHAT>?
>
> <HOW MANY> *X* </HOW MANY> <WHO> *African-American, female* </WHO> <WHERE> *New York City* </WHERE> <WHEN> *01/1999-12/1999* </WHEN> <WHAT> *Unemployment* </WHAT>?

Results and Future Work

While a fully-implemented search engine which utilizes the statistical-query grammar was not developed as part of this funded project, the results we can present reflect the ability of the grammar to provide coverage of new queries. A simulation of the query grammar on a small sample of new user queries, showed that 95% of the queries could be covered and covered accurately by the statistical-query grammar we developed.

We have several lines of development and testing we would like to pursue with this research. First of these would be a full implementation of the statistical search engine for testing on a larger sample of queries and on a set of tables whose labels for table, column, row, sub-row, and cell are represented in meta data elements accompanying the tables.

Secondly, we would like to extend the query grammar to the six classes of queries which we did not include in the sample on which the grammar was developed. The excluded classes of queries were not requests for which a statistic, per se, would suffice as an answer, e.g. they were requests for some action to be taken by the agency, questions about the manipulation of data, the availability of certain data, details of data collection, or what we labeled 'vague requests'. This last class represents those queries which would most likely need to be mediated by a human reference person, as they require extensive clarifying dialogue which currently is beyond the abilities of Natural Language Processing. As many of us believe, the digital reference services of the future will be a combination of automatic and human responses, with a system performing the initial triage.

Thirdly, it would be an interesting experiment to turn the dimensions we found in users' queries into templates and present these fill-in-the-black templates as a means for guiding users in how to formulate statistical queries. The two approaches to querying – NLP queries, and template-filled queries - could then be empirically compared as evidence as to which type of querying interface produces the best results and is easiest for information seekers, particularly those who are not expert in statistical information. To do this, we can see translating the journalistic elements into more recognizable aspects. Using this terminology, the system might structure templates on these more understandable labels:

- Population
- Location
- Time period
- Condition
- Quantification

In conclusion, we appreciate the support we have received from various government agencies who are currently striving to provide the best in information and services to the masses of citizens who are not familiar with the intricacies of statistical information. We believe that the research conducted in this project has served to provide useful evidence to both system developers and providers of statistical information on how a portion of naïve users' queries might be dealt with automatically, thereby allowing statistical experts to focus on those queries which require their specialized attention.

References

Haas, S. W. & Hert, C.A. (2000). Terminology development and organization in multi-community environments: The case of statistical information. In Proceedings of The SIG/CR Workshop on Classification.

Hert, C. (1999). What we know about users of statistical information. http://istweb.syr.edu/~tables/bground.htm

Hert, C. (2001). Studies of metadata creation and usage. In Proceedings of the Federal Committee on Statistical Methodology Research Conference.

Hert, C., Marchionini, G., Liddy, E.D. & Shneiderman, B. (2000). Interacting with Tabular Data through the World Wide Web. FCSM Statistical Policy Seminar. http://istweb.syr.edu/~tables

Liddy, E.D. (1998). Enhanced text retrieval using Natural Language Processing. Bulletin of the American Society for Information Science. Vol. 24, No. 4. http://www.asis.org/Bulletin/Apr-98/liddy.html

Liddy, E. D. (1991). The discourse-level structure of empirical abstracts: An exploratory study. Information Processing and Management, 27:1, pp. 55-81.

Liddy, E.D. & Liddy, J.H. (forthcoming). Developing a lexico-syntactic, semantic sublanguage grammar for statistical-queries.

Liddy, E.D., McVearry, K., Paik, W., Yu, E.S. & McKenna, M. (1993). Development, implementation & testing of a discourse model for newspaper texts. In Proceedings of the ARPA Workshop on Human Language Technology, Princeton, NJ.

Marchionini, G., Hert, C., Liddy, E.D., Shneiderman, B. (2000). Extending understanding of federal statistics in tables. Proceedings of the ACM Conference on Universal Usability. ACM, NY; 132-138.

Marchionini, G. (1999). Interfaces for understanding and using statistical tables. Proceedings of CHI 99. http://cpcug.org/user/hamilev/chi99/marchionini.htm

Sager, N. (1981). Natural Language Information Processing: A Computer Grammar of English and Its Applications. Addison-Wesley, Reading, Mass.

Sager, N., Friedman, C., Lyman, M.S., MD. (1987). Medical Language Processing: Computer Management of Narrative Data. Addison-Wesley, Reading, MA.

Tanin, E. & Shneiderman, B., (2001). Exploration of Large Online Data Tables Using Generalized Query Previews, Univ. of Maryland Computer Science Technical Report.