CERIAS Tech Report 2005-32

METHODOLOGY AND TOOLS FOR ONTOLOGICAL SEMANTIC ACQUISITION

by John M. Spartz, Evguenia Malaia, Courtney Falk

Center for Education and Research in
Information Assurance and Security,
Purdue University, West Lafayette, IN 47907-2086

Methodology and Tools for Ontological Semantic Acquisition

John M. Spartz
English Linguistics
Purdue University
West Lafayette, Indiana 47907
jspartz@purdue.edu

Evguenia Malaia
CERIAS & NLP Lab
Purdue University
West Lafayette, Indiana 47907
emalaya@purdue.edu

Courtney Falk
CERIAS
Purdue University
West Lafayette, Indiana 47907
court@cerias.purdue.edu

Abstract

This article focuses on the most important facets of ontological semantics, and more specifically, on the process of ontological semantic acquisition for linguistic students and researchers inexperienced in this emerging field, regardless of their prior work in computational linguistics, NLP, or lexical semantics. The overarching goal of this text is to provide interested parties with a synthesis—a guide to the methodology and tools—so they might more efficaciously continue the work that has begun in ontological semantics at the Computing Research Laboratory of New Mexico State University, the University of Maryland, Baltimore County, and, most specifically, Purdue University.

1 Introduction

According to the seminal work in the field, Ontological Semantics, ontological semantics is "a theory of meaning in natural language and an approach to natural language processing (NLP) which uses a constructed world model, or ontology, as the central resource for extracting and representing meaning of natural language texts, reasoning about knowledge derived from texts, as well as generating natural language texts based on representations of their meaning" (Raskin and Nirenburg, 2004). It is this definition that facilitates a clearly defined discussion of ontological acquisition, the process of building that "constructed world view," that ontology. The general features of ontological semantics as a whole and the justification of the approach are dealt with in "Ontological Semantic Support for a Specific Domain" (Raskin, et al., 2005); the chief objective of this article is to provide a clear guide to the tools and methodology for acquiring and building ontology.

Acquisition is the lifeblood of ontological semantics. Through the acquisition process, trained acquirers describe the ontological backbone to this natural language processing approach. But, the acquisition process can be difficult, redundant, and extremely time-consuming, leading to a variety of errors and an ultimate slow-down of an already lengthy and difficult human effort. Proper preparation for acquisition is paramount to success; gathering and properly utilizing the available acquisition tools is a compulsory step in pre-acquisition practice.

2 Pre-Acquisition Tools

When approaching ontological acquisition, novices need to be equipped with a variety of tools to delve into the acquisition of a new domain, concept, or lexical item for the ontology. It is these tools, these areas of ontological semantics, that will function to turn "novices" into "masters," or at least help them along the continuum.

One of the first, and it might be argued, most important tools for future acquirers is a clearly stated set of terms and accompanying definitions relevant to acquisition. Some of those necessary terms and rudimentary definitions include the following standard language in ontological acquisition:

- **Avoider (KBAE):** a stand-alone, offline program that allows for simple traversing of the ontology.

- **Concept:** an ontological item in an ontological tree. It must be unique and is preferably maximally specified. Its name does not constitute its meaning and functions as a parent to a host of lexical items in the ontology.
- **Child**: a concept (in an ontological tree structure) inheriting some features from the concepts from which it descends. A further specification of a parent concept.
- **Corpus**: a large collection of written, and sometimes spoken, examples of the usage of a language, used for linguistic analysis.
- **Domain:** a more or less concretely specified field of knowledge to which ontological semantic processing is applied.
- **Fact Database:** a storage of instantiations of events and objects, which is used for text processing and assists in disambiguation. A numbered collection of instances of generic ontological concepts.
- **KBAE:** Knowledge Based Acquisition Editor; a program used to expand static sources in ontological semantics, most notably in the ontology and lexicon. It provides information about inheritance, allowing viewing of concepts and lexical items and their relationships.
- **Lexicon:** the group of words available to the language processing system within an ontology. Each concept consists of numerous members of the lexicon.
- **Onomasticon:** a depository of proper names used to complement ontology. It is a token database for text processing, which groups all instances into one of four categories: animate, organization, time-period, and geographical entity.
- **Ontology:** an inventory of concepts related to the "real world." A detailed and constructed world model, containing information about 1) the physical world, 2) discourse participants, and 3) the communicative situation. Its main function is to provide conceptual, categorical description of the real world, so as to make possible a description of lexical items in a consistent and logical manner.
- **Polysemy:** the ambiguity of an individual word or phrase that can be used in different contexts to express two or more different meanings.
- **Parser:** Fundamental part of the dynamic text-processing algorithm that leads to the creation of TMR. Turns text into TMR, or computational representation of its meaning.
- **Semantic Analyzer:** Fundamental part of the dynamic text-processing algorithm that leads to the creation of TMR. It carries out the tasks of establishing propositional dependencies and deals with the pragmatic aspects of text: style, speaker attitude and goals, etc.
- **Text Analyzer:** defines the chief meaning of a proposition or multiple propositions in a sentence, resolves lexical ambiguities, from the existing text, generates the TMR.
- **TMR:** Text Meaning Representation; the output of the parser and a computational description of the text's overall semantics
- **Tokenizer:** Fundamental part of the dynamic text-processing algorithm that leads to the creation of TMR. It breaks text into usable strings, dealing with special characters, numbers, symbols, punctuation, and all other "ecological" issues.

Ultimately, it is these terms and their accompanying definitions that will assist linguists new to this area of study in understanding and applying that understanding to actual acquiring.

Yet another, but extremely important tool for linguists hoping to successfully acquire concepts and lexical items in any particular domain, is a dictionary—more specifically, a dictionary specific to the domain area. For example, when acquiring in the medical domain, researchers should use a medical dictionary, in the legal domain, a law dictionary, etc. Dictionaries are not only useful in providing definitions for humans in the ontology, housed on a centralized acquisition tool such as Purdue University's KBAE, but also in polysemy reduction, one of the major areas of focus for master acquirers. Yet another benefit of a solid, domain-related dictionary is the possibility of "rapid propagation" (Raskin and Nirenburg, 2004), wherein skilled acquirers can quickly use an already acquired lexical item's formalism to obtain numerous related items.

3 Acquisition Tools

In order to begin the actual process of acquiring ontological items, it is most efficacious to have access to some sort of automated acquisition tools. The technology for acquiring and housing the acquired ontologies, while existing, is somewhat lacking in quality. One of the earliest efforts is known as the Knowledge Base Acquisition Editor (KBAE) (See Appendix A, Figure 1), a centralized acquisition tool with a web-based interface. This piece of software was developed under the supervision of the Computing Research Laboratory of New Mexico State University. The existing KBAE program is run on a remote server, which also runs a web-hosting program, allowing acquirers to log in with a secure name and password and edit the knowledge base. The strengths of KBAE are 1) its web-based interface, which allows all users with a web browser to acquire, 2) its centralized nature, which eliminates merging of multiple copies of the ontology, and 3) the ability to standardize the structure of ontological entries.

Unfortunately, KBAE has some serious shortcomings, which have, ultimately, led to its disuse. The amount of time KBAE takes to process acquisition or editing actions can, at times, take upwards of a minute and a half. This is an unacceptably long period of time in acquisition, where tools are developed to accelerate the process. The swiftness problem is a product of KBAE's lack of scalability; the cause of the problem is KBAE's structure, which is, in part, comprised of its own database executable, instead of a third-party software solution, which would allow for future expansion and refinement.

After KBAE was abandoned, tool development shifted to stand-alone applications with smaller sets of functionality. Steve Beale of the Institute of Language and Information Technologies at the University of Maryland at Baltimore County developed a tool appropriately known as KBAE Avoider. Avoider, because of its structure, allows an acquirer to browse the ontology while acquiring much more quickly than the original KBAE tool.

Research and development of acquisition tools is ongoing. On the heels of KBAE Avoider, Courtney Falk of Purdue University developed ChangeTool (See Appendix A, Figure 2) for use in Purdue University's CERIAS and NLP labs. The new tool was created as an application that would allow entries to be created or edited according to the structure defined in the aforementioned *Ontological Semantics* (Raskin and Nirenburg, 2004). It was also created so as to avoid the duplication of functionality already provided by KBAE Avoider.

Both stand-alone tools, KBAE Avoider and ChangeTool, utilize GUI interfaces to improve the ease of acquisition, but each is written in a different programming language. KBAE Avoider uses Lisp, which suits the parenthetical format of the ontology and lexicons quite well; unfortunately, Lisp is difficult to compile, and once compiled, it can't be easily run on different versions of Windows, let alone other operating systems. ChangeTool uses the Java language, which, while requiring the Java virtual machine to be installed on the acquirer's computer, also allows the program to be run on any computer, once it is initially compiled.

The preeminent, apparent benefit of a centralized acquisition tool is that it avoids platform dependency issues that affect KBAE Avoider and, to a lesser extent, ChangeTool. Centralized acquisition has multiple other benefits, such as the ability for administrators (master acquirers) or project heads to adjudicate conflicts between lower level acquirers. Keeping development copies of the ontology and lexicons also means rapid updates to production data, since separate files don't need to be gathered and then combined. Ultimately, automated tools such as KBAE, Avoider, and ChangeTool allow acquirers access to and functionality of the existing and building ontologies not provided by the original KBAE tool.

In all, it is the combination of the pre-acquisition and acquisition tools that makes for successful ontological semantic acquisition. Although, it must be noted that it is not without a clearly defined methodology that an amateur acquirer can begin the daunting task.

4 Methodology

One "tool" paramount to success in ontological semantic acquisition is a clearly defined methodology for that process. In order to most effectively illustrate the methodology, the processs description will be focused on acquiring for a domain, specifically in the area of Digital Identity Management (DIM), one of the current domains of interest for researchers at CERIAS at Purdue University

In order to comprehensively survey a domain, it is necessary to divide it into several "subdomains" and identify any major influences for each, using them as sources for corpora in each specific subdomain. For example, in the DIM domain, the following subdomains were determined: 1) social aspects of identity management, 2) technical architectures proposed fro identity management by various sources, 3) psychological issues arising from the uses of digital identities, legal frameworks governing the use of digital identities (mostly implemented as laws pursuant to personal data handling), and 4) biometrics as an emerging and controversial field, allowing, in many cases, unique identification, but also prone to problems specific for the field.

One of the problems an acquirer faces in many rapidly developing domain fields (such as DIM) is the need to distinguish emerging *concepts* and cross-applicable vocabulary from the *lexical items* invented ad-hoc by vendors and researchers, which will not be used by anyone else. Thus, it is extremely important to keep track of the source of the particular corpora. For example, in DIM, the rhetoric used in the corporate world is vastly different from that used by government organizations, and that is still dissimilar from the rhetoric of not-for-profit organizations, international organizations, and academic researchers. The difference among these entities is not simply the terminology, but rather the attention to particular aspects of the actual transactions involved in Digital Identity Management.

From the linguistic standpoint, however, the interest lies in accurate semantic and world-view-information descriptions of all terminology pertaining to the domain (which includes the ontological support hierarchy and "peripheral" terms, which, even if not overtly present in a particular text, surface in the semantic description of the domain). For the aforementioned reasons, it is necessary to construct a domain "topic-source variability matrix," which deals with all aspects of DIM, as is advisable, regardless of the domain. Table 2 (See Appendix B) is an example of such a topic source matrix; the top row (sources) consists of the agents involved in a particular topic discussion. The leftmost column represents the aforementioned subdomains. By filling out the entire matrix and working with the corpora represented in this heuristic, external validity of the corpus is ensured.

4.1 Corpus-Based Pre-Acquisition Methodology

Once the corpus is determined and the validity is ensured, some actual acquisition can begin. The main question that often must be solved during the acquisition process is how to delimit the boundary between ontological and lexical items, which need acquiring. Acquirers must initially ask themselves the question: Is an item sufficiently different conceptually to be introduced as an ontological item, or can it be kept to the domain of the lexicon? For the purposes of the present work in ontological acquisition, this question, pertaining to the parsimony of the ontology, is solved on the basis of the following criteria:

1. The introduction of the new concept is justified if it can be used for lexical items other than the one in question.
2. Introduction of the new concept is justified if it can be used for other ontological items.
3. The grain size of the semantic description that the present system is aiming for is smaller than the current ontological description and further specification is needed, therefore, introduction of the new concept is justified by reasons of granularity of semantic description.
4. Alternative methods of semantic description of a lexical item involve concepts that are also not members of the ontology and need to be added.
5. Criterion 1 and 2 (general applicability of the concept to ontological and lexical descriptions) are in favor of the previously considered concept.

If the previously delineated criteria are efficiently applied and resolved in favor of the concept, that concept is then added to the ontology. If not, the lexical item in question is specified through the already-available concepts in the ontology.

Before acquisition of ontology and lexical entries for the domain can begin, it is necessary to create the basic structure for the ontological subtrees and determine the lexical items that need to be added. Table 1 contains the action-plan table for pre-acquisition methodology.

The top-down acquisition involves, first of all, adding new properties to the ontology. The property list is the one that, on one hand, allows for a rigorous description of concepts in the ontology, but on the other hand has to be limited in size for the purposes of thorough description. The acquisition of properties is driven by both the question of grain size for the ontological description and the need for deep semantics in the description of lexical and ontological entries. However, it is advantageous for the list to be limited, both for the purpose of non-proliferating limited-use concepts, and for the ease of any future acquisition effort. Thus, the list of necessary attributes and relations needed for the description of a domain is the first one on which a decision has to be made.

Top-Down Methodology	**Bottom-Up Methodology**
Delimit **the corpus**, dealing with all the aspects of the particular domain (topic-source matrix). Split it in two parts for validity check if possible.	Run an item from each square of the source matrix **corpus through the available lexicon** and filter out lexemes that are not yet available.
Map out an ontological tree for the most important concepts for each subdomain; establish the **necessary properties** for the overall domain and acquire those that are not already in the existing ontology.	**Sort the lexical items** determining whether they belong to the particular domain.
Create ontological **subhierarchies** needed to support the subdomains.	
Decide on **multi-word expressions** (phrasals) necessary for the domain specific vocabulary.	**Acquire** lexical items. Add non-domain lexical items to "IOU"/common word-stock list (also used for running the corpus through).
Check for **multiple meanings** of available items in the lexicon, so that the senses in the particular domain are represented.	If necessary, expand the corpus (2 or 3 items from each source-topic) for another **validity check**.
Result: ontological hierarchy and lexicon for the specific domain.	

Table 1. Approaches to domain acquisition

The final step in ontological acquisition is to check on whether all necessary meanings of lexical items are represented in the items already in the lexicon. The filtering program used for corpus-based pre-acquisition is not intended for the creation of TMRs, so in the absence of an analyzer, it is necessary to verify that each lexical item in the lexicon has its meanings for the domain listed in its semantic description, and it is tied to the ontological concepts necessary for the domain. This step can be done at any time, even after corpus-based (bottom-up) acquisition, and it concludes the top-down process of domain acquisition.

In order to extract the lexical items from the corpus, we wrote a small program that runs the corpus (one article at a time) through the already-existing lexicon. It also does minimal morphological analysis, eliminating some of the morphological forms of existing words. The main purpose of the program is to keep track of already-acquired lexical items.

The output of the program is a file with all the words that are not found in the main lexicon or "common words" file. The "Common words" file at the moment contains: all contractions (isn't, it's, etc.), tensed irregular verbs, pronouns, conjunctions and other closed-class lexical items that will eventually be processed by the analyzer and contribute to TMR. For the purposes of the present work, we keep them filtered out, since we aim to acquire the vocabulary and ontology for the domain.

The lexical items from the output file are consecutively sorted into "domain" and "non-domain" items. It is a flexible division, based on the following criteria:

1. Can the lexical item be ontologically described using the properties added for the specified domain?
2. Is it conceptually related (at least in one of its senses) to the lexical items already acquired?
3. Does it appear more than once in the articles pertaining to the domain?

4. Does it contain semantic information belonging, fundamentally, to TMR (e.g. irregular tensed verbs or some adverbs), in which case, the lexical item has to be processed by the analyzer and does not have to be part of the lexicon?

If the answers to the first three questions are positive, the item is resolved as belonging to the domain and is acquired, unless the answer to the last question is also "yes". In all other cases, the item is relegated to the "common words" file, and consequently "filtered out" of the corpus.

The last step of the corpus-based approach is the validity check for the domain. For the purposes of the present work, we rely mainly on the topic-source matrix to provide external validity for domain coverage. This strategy is instrumental in acquisition, regardless of domain. These basic steps in building an ontology are fairly straightforward, once clearly defined and explained to a new builder. In all, the methodology can be synthesized in the following manner:

1. Acquire domain knowledge.
2. Organize the ontology.
3. Flesh out the ontology.
4. Check the work.
5. Commit the ontology (Denny, 2002).

In all, there is much to know about ontological semantics and ontological acquisition. It is essentially the aim of this article to give linguists either just beginning or unfamiliar with ontological acquisition some guidance in the quest to become proficient in the field.

References

Steve Beale. *KBAE-Avoider*. 2002. Retrieved January 29, 2005.

Computing Research Lab. 2002. *Knowledge Base Acquisition Editor*. Retrieved March 1, 2005.

Michael Denny. "Building: A Survey of Editing Tools." 2002. Accessed online at http://www.xml.com/lpt/a/2002/11/06/ontologies.html

Courtney Falk. 2005. *ChangeTool*. Retrieved February 14, 2005.

Victor Raskin, and Sergei Nirenburg. 2004. *Ontological Semantics*. MIT Press, Cambridge, MA.

Victor Raskin, Katrina Triezenberg, Evguenia Malaia, and Olga Krachina. 2005. Ontological Semantic Support for a Specific Domain. In this volume.

Appendix A. Tools for Ontological Semantic Acquisition

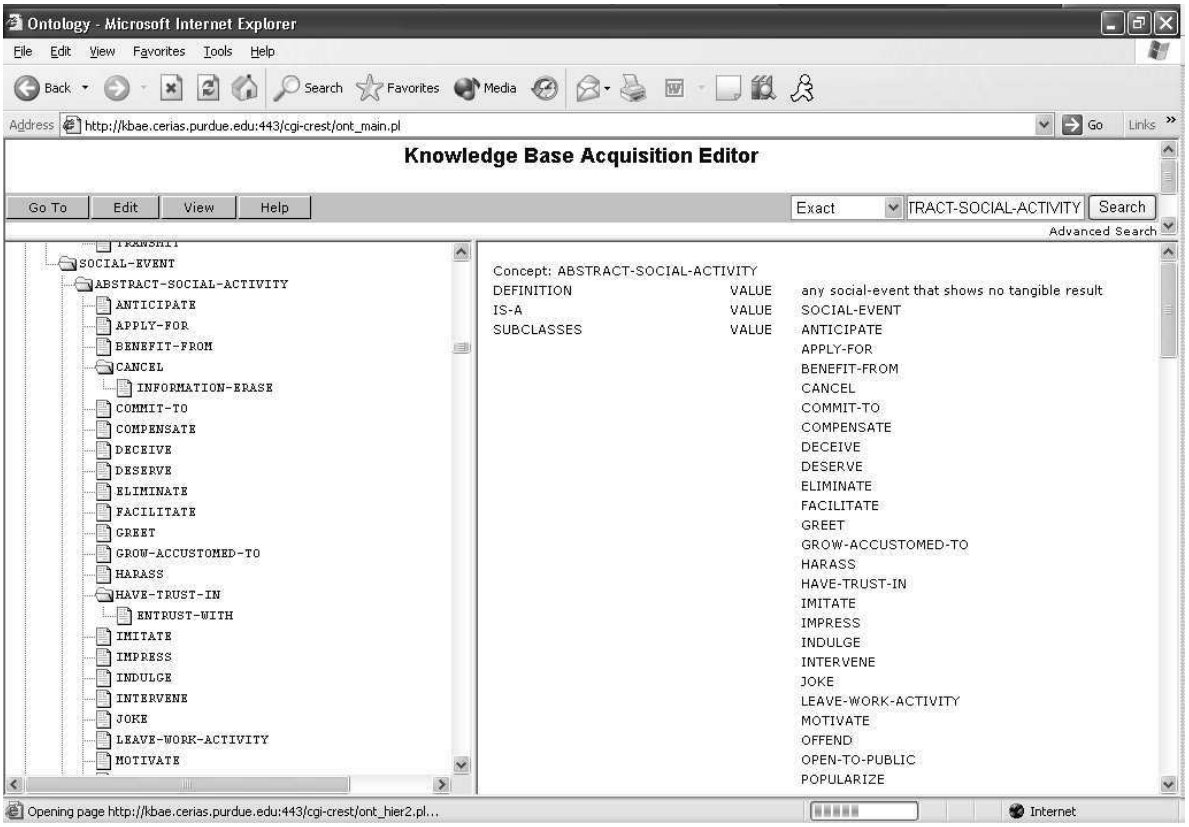

Figure 1. KBAE view of ABSTRACT-SOCIAL-ACTIVITY

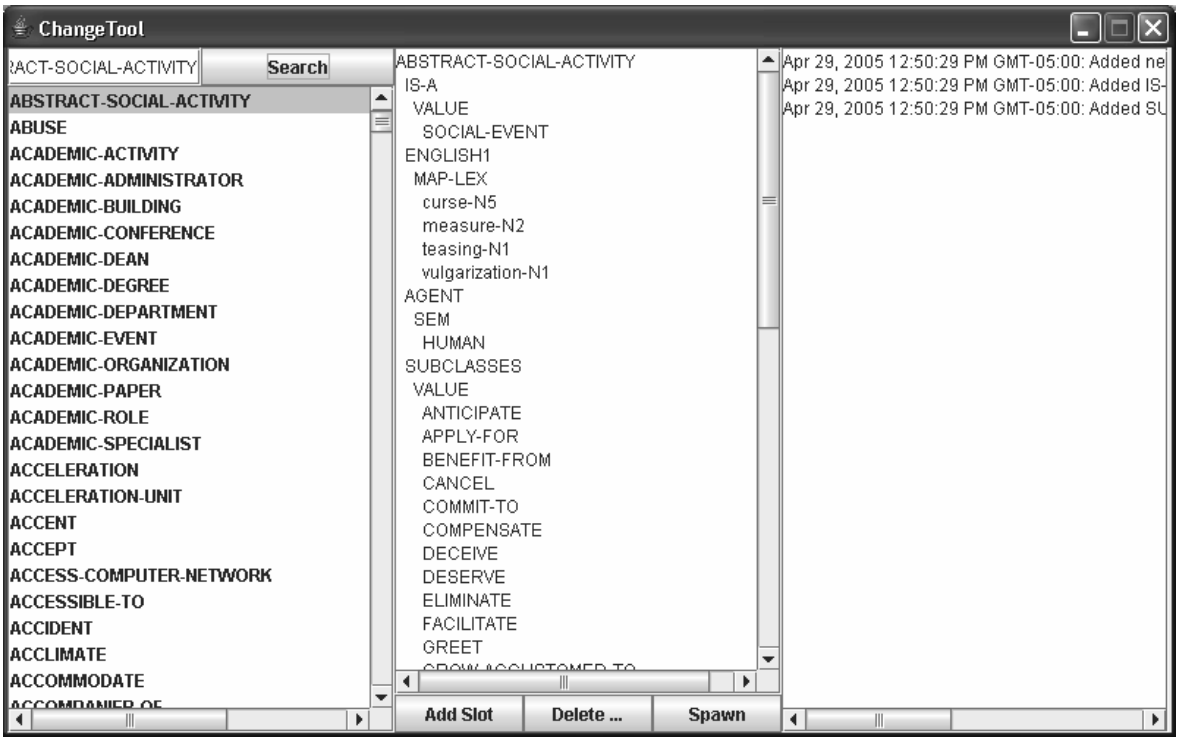

Figure 2. ChangeTool view of ABSTRACT-SOCIAL-ACTIVITY

Appendix B. Methodology for Ontological Semantic Acquisition

Source / Topic	Non-profit organizations	Business and industry groups	U.S. federal agencies	International organizations	Academic research
Biometrics and its usage	Biometrics Consortium; EFF; National Biometric Security Project	Precise Biometrics; BioPassword; Cognitech; Florentis	US Dept. of State; Biometrics Consortium	OECD; Council of Europe	International Biometric Society; US NBTC
Psychology of digital identity deployment	Identity Theft Resource Center	IBM	NA	NA	Sherry Turkle
Technical implementation of identity management schemes	W3C; Association for Automatic Identification and Data Capture Technologies	Liberty Alliance; IBM; RSA Security; Motorola; VeriSign		OECD; IEEE	Mike Atallah, http://xxx.lanl.gov/archive/cs
Economic viability of IM schemes	TRUSTe	MS; Siemens; Applied Digital	Federal Trade Commission	OECD	Mills
Social aspects of using various DI schemes	Consumer Professionals for Social Responsibility; EFF; Electronic Privacy Information Center	Liberty Alliance; Microsoft	NA	United Nations; The Global Internet Liberty Campaign	Howard Sypher
Legal aspects of DI use	American Civil Liberties Union; EFF	RSA Security (advisory)	US Department of State; FTC	United Nations	Lawrence Lessig

Table 2. Corpus-based approach to lexicon acquisition: matrix for variability of sources in example domain