Inducing Ontologies from Folksonomies using Natural Language Understanding

November 13th 2009

Sponsored by

Defense Advanced Research Projects Agency (DoD)
(Controlling DARPA Office)

ARPA Order AW78-00
SBIR SB082-032 – Phase I

Issued by U.S. Army Aviation and Missile Command Under
Contract No. *W31P4Q-09-C-0386*

Name of contractor: **Lymba Corporation**
Principal investigator: Dr. Dan Moldovan
Business address: 1701 N. Collins Blvd., Suite 3000, Richardson, TX, 75080
Phone number: (972) 680-0800

Effective date of contract: April 28th, 2009
Short title of work:

Contract expiration date: December 31st, 2009
Reporting period: April 28th – November 13th, 2009

1. Introduction

Social bookmarking is rapidly emerging as a tool for users to associate subjective descriptions (tags) to web pages, which help them organize and recall information of interest. Sharing bookmarks enables the discovery of users with common interests, resources with common tags as well as the generation of folksonomy – a flat, user-created lexicon of terms that the community adopts to associate with resources of interest. By explicitly capturing and representing tag semantics in a taxonomy or ontology, the information structure of user tags is revealed, thus, facilitating machine understanding of user interests. Furthermore, ontologies induced from folksonomies allow users to visualize and navigate through the information structures in the tag space and to discover semantic relations between tags.

Advanced linguistic processing of tags results in a better organization and management of folksonomies as well as improved sharing of resources. As highlighted by Golder and Huberman [4], the users' uncontrolled vocabulary includes different types of variations and ambiguity, for instance, case sensitivity of tags, use of space or punctuation as delimiters, both singular and plural forms, same tag applied in different context, and synonymy of concepts. Adam Mathes[1] notes *The sheer multiplicity of terms and vocabularies may overwhelm the content with noisy metadata that is not useful or relevant to a user.*

In order to overcome these problems, Lymba Corporation employed its Natural Language Processing (NLP) and automatic ontology generation technologies to process and induce ontologies from folksonomies and to develop applications that exploit this latent structure of folksonomic tags.

This report describes our technical objectives for the SBIR SB082-032 "Inducing Ontologies from Folksonomies using Natural Language Understanding" project (Section 2). In Section 3, we detail our Phase I technical accomplishments, the various automatic procedures that we implemented into a prototype system that creates an ontology from a given input folksonomy. Our implementation follows the proposed research tasks and shows not only the feasibility of our approach, but also the linguistic tools and resources needed to build a system that exposes semantic structures of folksonomies.

We describe our plans for Phase II of the project in an enclosed document.

2. Technical Objectives

During the Phase I period, Lymba studied the feasibility of an automated system that exposes a folksonomy's semantic structure. We developed a prototype system that implements our proposed design and exposes a folksonomy's semantic structure by building an ontology of tags which can be improve various folksonomy-related applications, such as automatic generation of user, document, or tag recommendations or collaborative tagging across multiple social

[1] http://adammathes.com/academic/computer-mediated-communication/folksonomies.html

bookmarking applications in addition to a folksonomy visualization, browsing and search application.

As part of this effort, we identified the desired characteristics of a representation that captures tag semantics as well as enables discovery of semantically related tags in the folksonomy. The formal semantic representation of the folksonomy links the semantics of tags back to concepts in an underlying ontology (WordNet).

In order to derive a rich semantic representation of the folksonomic tags, Lymba developed mechanisms to normalize the lexical, syntactic, and semantic variations present in the folksonomic data. For this purpose, we exploited not only a tag's textual information, but also its associations with other tags and with documents as created by users as part of the social bookmarking data. Lymba has the unique capabilities to automatically process the content of bookmarked documents and accurately understand a tag's meaning based on its associations.

Once each tag's meaning was captured in a rich semantic representation, Lymba identified a series of classification procedures that produce numerous tag-tag relationships that complete the ontology induced from the flat lexicon of folksonomic tags. SYNONYMY, ISA, PART_WHOLE, SIMILARITY, DOMAIN, ATTRIBUTE, and other relations between tags expose the folksonomy's ontological organization. We note that Lymba has previously built domain-specific ontologies using its automatic ontology generation module.

The rich semantic network of folksonomic tags provides new dimensions for exploring the social bookmarking data. We explored different methods for enhancing information discovery and access applications that use both the semantic and social networks of social bookmarking data. Lymba has developed a resource discovery engine that processes social bookmarking data, models users, and recommends documents that would be of interest to the user based on the information available in the tag space. However, using only few tag normalization techniques, our current approach does not exploit a full ontological structure of the folksonomy that would provide new enhancements of our existing system.

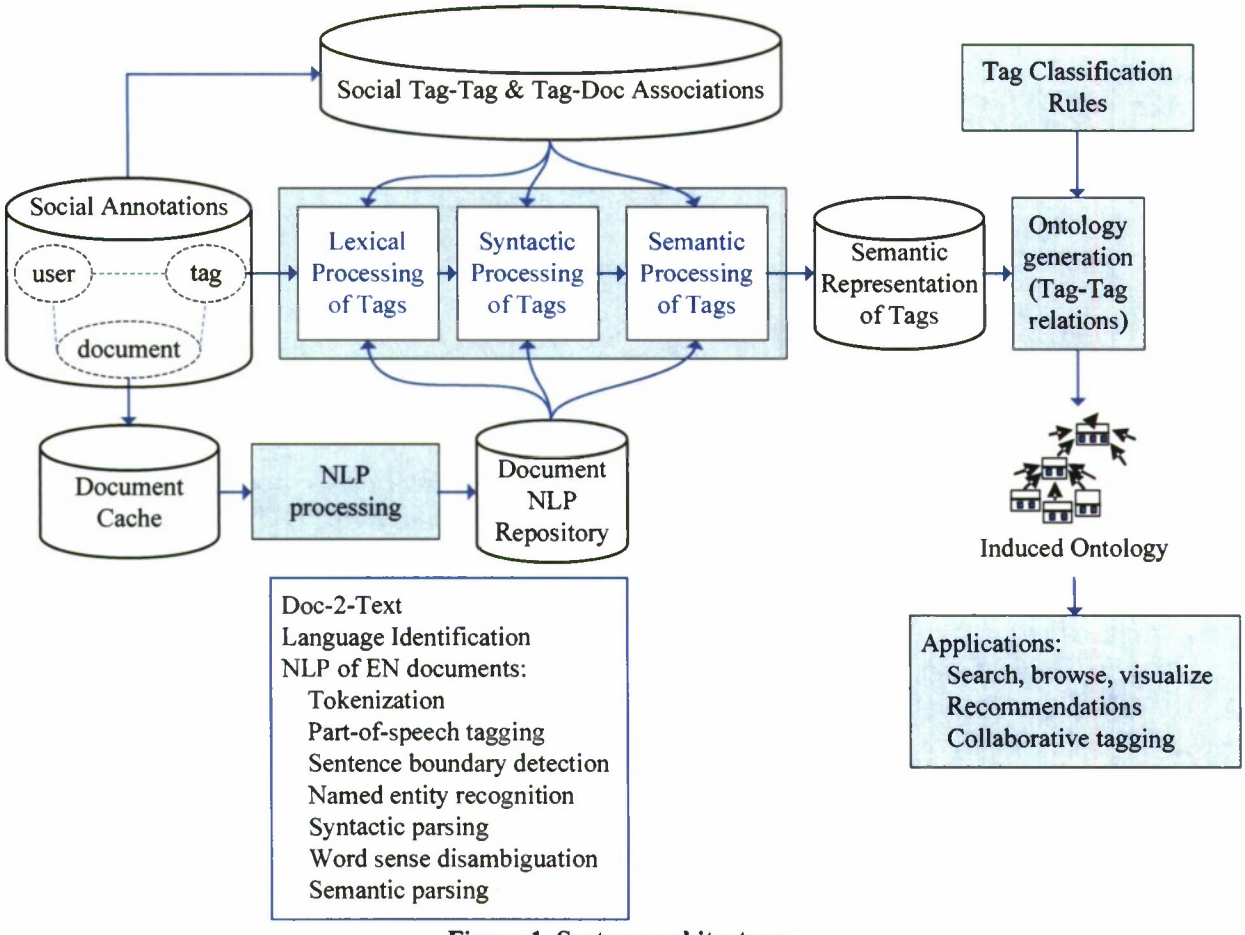


Figure 1. System architecture

In Figure 1, we show the architecture of our prototype system, which implements the technical objectives outlined above.

3. Technical Accomplishments – Prototype Description

Our efforts for the past six months were focused on building a prototype system that derives the semantic structure of an input folksonomy. This initial version of the system implements our proposed research for both the important step of understanding the tag semantics as well as the derivation of tag-tag semantic relations that expose a semantic structure within the flat folksonomy. We plan to continue the development of this system in Phase II by (1) introducing new sources of information in the tag understanding process, (2) expanding the processing to languages other than English, (3) making the system scalable to real world size datasets, (4) converting its current batch mode operation to a live real-time analyzer, and (5) developing applications that exploit the ontological structure of the folksonomy.

For the development and testing of this prototype system, we used the social bookmarking dataset described in Section 3.1.

3.1. Experimental Data

For Phase I, we used social bookmarking data collected from the del.icio.us bookmarking service (http://delicious.com) that allows users to tag, save, manage, and share web pages from a centralized source.

Our initial dataset, as described in Table 1, is stored in a MySQL database and can be browsed using the Scuttle social bookmarking tool (http://sourceforge.net/projects/scuttle). It includes all (user, document, tag) social annotations stored publicly in del.icio.us between May 19th and June 4th.

Table 1. Social bookmarking data collected from Delicious between May 19th and June 4th

(user, document, tag) triplets	7,162,536
(user, document) pairs	2,362,794
unique number of users	359,000
unique number of documents	975,673
unique number of tags	342,314
average tag / (user, document)	3.03
average tag / document	7.34
average document / user	6.58

We aimed to having a dataset that contained a substantial set of tags, which will ensure a good representation of the folksonomy they generate by allowing us to rely on the content of the labeled documents and other co-occurring tags in the tag disambiguation process. We note that we downloaded the bookmarked URLs and processed the cached documents using Lymba's suite of NLP tools focusing mostly on English textual documents whose content will be used during the process of understanding each tag's semantics. 2.66% of the documents were not reachable by our crawling tool. 4.21% of the remaining documents have non-textual content (images, audio or video files, etc.). 23.03% of the remaining textual documents have non-English content.

Our initial processing and testing was performed on this large real-world social bookmarking dataset. However, since our goal for Phase I was to show the feasibility of our approach, all our future experiments were performed on a smaller dataset, described in Table 2, dataset created from the social bookmarking information collected from Delicious. We plan to use the original dataset (Table 1) for the 2nd Phase of this project.

Table 2. Phase I experimental dataset

(user, document, tag) triplets	148,709
(user, document) pairs	113,313
unique number of users	58,198
unique number of documents	83,827
unique number of tags	8,460
average tag / (user, document)	1.31

average tag / document	1.77
average document / user	1.94

For this dataset, we started with a set of 8,460 tags and added to our social bookmarking annotation set up to 25 bookmarks – (user, document) pairs – for each tag included in the dataset. This ensures that each tag is well represented in terms of associated document content. This dataset was used to develop and test our proposed tag understanding procedures (Section 3.3) and our proposed tag classification rules (Section 3.5).

3.2. Defining a Semantic Structure for Folksonomies

Folksonomies are collections of tags. Thus, our initial efforts in designing a formal representation of folksonomies focused on the tags and their representation. For each tag, Lymba created a rich semantic representation that captures the concepts mentioned in the tag text and their semantic relations. Therefore, each tag becomes a rich semantic graph that can be easily exploited during the process of organizing the tags (transforming the folksonomy into an ontology). We show several examples in Figure 2

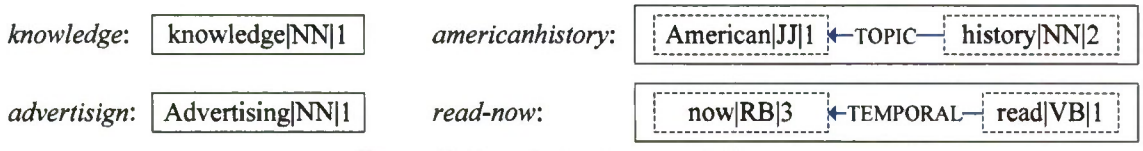


Figure 2. Sample tag representations

We note that each concept part of a tag representation is linked to its corresponding WordNet synset. For example, *american|JJ|1* is part of synset id *02927512*. These links enable the system to identify synonyms ((word, sense) pairs that denote the same concept in WordNet, therefore, are part of the same synset). Each tag will be accompanied by certain metadata, which includes language information, bookmarking information, certain count/frequency statistics, etc. In Figure 3, we display the SYNONYMY cluster of {*cognition, knowledge*} with their corresponding links to the original folksonomic tags that map to this cluster and their associated social bookmarking data information (for each *tag*, we list several (*user*, *document*) pairs where *user* assigned *tag* to *document*). The SYNONYMY clusters groups a set of normalized tags – semantic representations of folksonomic tags – derived using the tag understanding procedures described in Section 3.2.

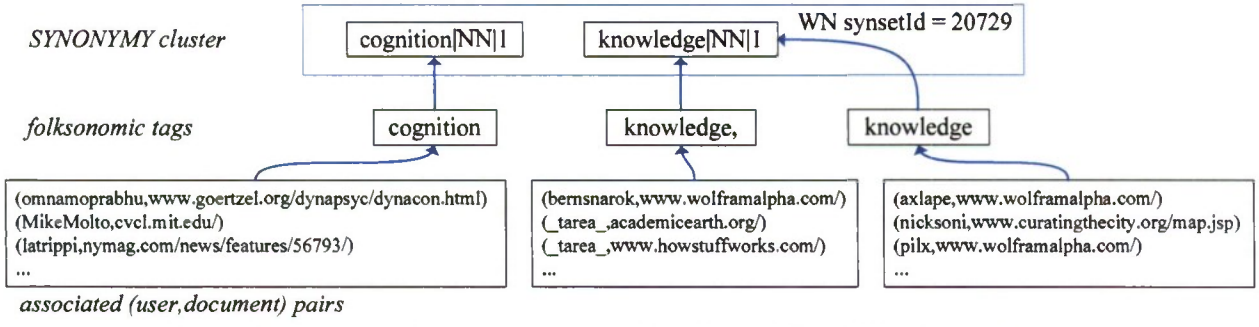


Figure 3. SYNONYMY cluster of normalized tags with social meta information

At the folksonomy level, semantic relations, such as ISA, PART_WHOLE (PW), SIMILARITY (SIM), etc. link the tags, inducing a rich semantic structure for the given folksonomy. Therefore, folksonomies are represented as rich semantic graphs whose links are the semantic relations that connect the tags forming the folksonomy, which constitute the nodes of the representation. For semantically equivalent tags, a single semantic representation is used and corresponding normalization links make explicit these tag connections.

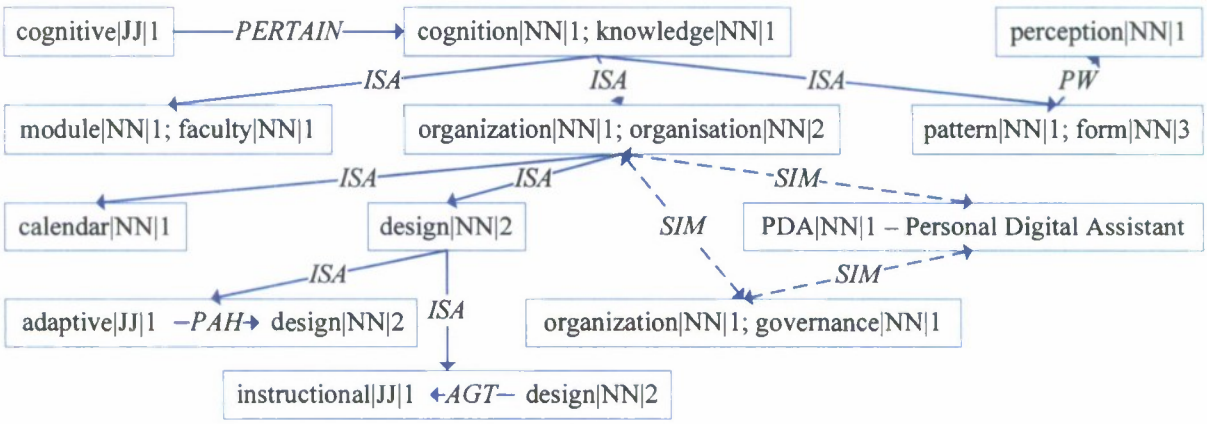


Figure 4. Sample ontology

In figure 4, we display a portion of a folksonomic structure. Its nodes are the SYNONYMY clusters detailed in Figure 3 (here, we display only the set of synonymous normalized tags). The links represent the semantic relations identified between the tags using the classification procedures described in Section 3.4.

3.3. Capturing and Representing Tag Semantics

Lymba broke down the process of understanding tags into eight different linguistic processing steps (Figure 5). Each stage uses three sources of information that provide complementary information to our prototype system: (1) *the tag space of the folksonomy*: the text of each tag is used to derive information about the tag, (2) *the social bookmarking data*: tag associations augment and refine the initial understanding of a given tag, and (3) *the content of textual documents*: situating a tag within the larger semantic context of the documents it was assigned enhances the existing understanding of a given tag.

The various linguistic processing steps we implemented as part of the process of capturing tag semantics can be classified as lexical, syntactic, and semantic.

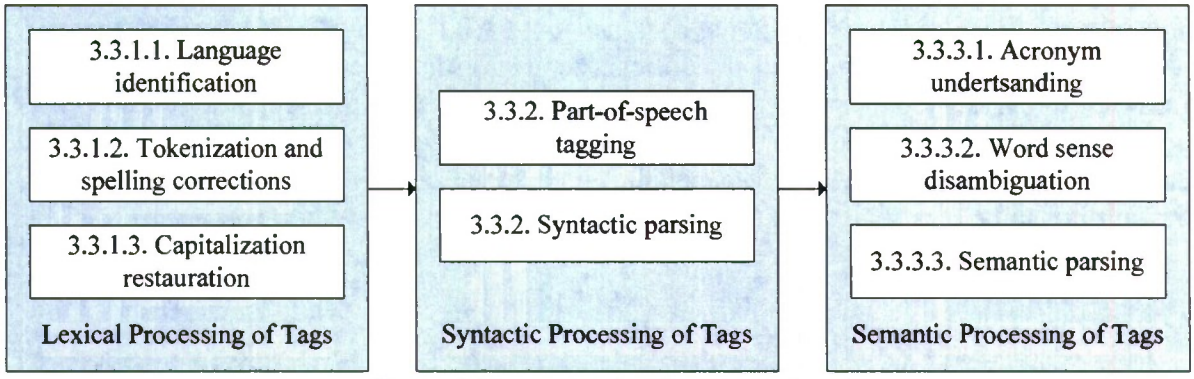


Figure 5. Tag understanding processing steps

3.3.1. Lexical understanding of tags

The lexical understanding of a tag includes the following stages: language identification, spelling corrections, tokenization and capitalization restoration.

3.3.1.1. Language identification

For the language identification step, we made use of Lymba's language identification module, which was expanded to include the 24 most frequent languages that we identified for our social bookmarking data[2]. These include Arabic, Chinese, Japanese, Korean, Russian, Turkish and many European languages. Each tag text was analyzed and matches against the various dictionaries were attempted. If a definite match was made, the tag's language was identified. If two or more languages had similar matching scores, the language of tag was decided based on the language of the documents that were labeled with that tag. We note that universal words, such as the numbers and most technical terms and names (e.g., linux, css, google), were tagged as belonging to the English language.

Evaluation

For our dataset of 8,460 tags, 91.65% of the tags were marked as belonging to the English language. Other most frequent languages include Spanish (1.85%), Portuguese (1.72%), and German (1.47%). We manually verified the correctness of language value attributed to 10% of the tags and the system is 97.87% accurate in assigning a language to a folksonomic tag. Because our dictionaries were built using Wikipedia in various languages, most errors occurred while assigning English as a language to non-English tags, which appear in English Wikipedia articles that contains many more entries when compared with other Wikipedia collections. Fewer errors are caused because the content of the document labeled with the given tag was not available for language analysis. Examples include *davidleeroth* and *vanhalen*, which are used to tag an all-flash website whose textual content was not derived by our system (http://www.thetyser.com).

[2] We collected a document's language information as part of the meta information we downloaded when we created our local cache for each bookmarked URL. A set of 34 languages were used to create the URLs' contents. Examples of the most frequent languages include British and American English (EN), German (GE), Spanish (ES), Japanese (JA), French (FR), Russian (RU), Italian (IT), Portuguese (PT). Examples of the least frequent languages include Tamil (TA), Breton (BR), Glacian (GL), Serbian (SR), Latvian (LV) and Irish (GA).

3.3.1.2. Tokenization and spelling corrections

The process of verifying whether a tag text belongs to a certain language vocabulary not only helps us identify the language of the tag, but also determine whether the tag is a single token. If a tag was found among the words that constitute the vocabulary of a language, it is a single token. If it was not found, then (1) the tag contained two or more words glued together – these should be tokenized/separated for a correct understanding of the tag semantics, or (2) the tag was indeed a single token, but it was spelled wrong by the social bookmarking user, or (3) a combination of (1) and (2). Therefore, for each unmatched tag, we generated a list of correctly spelled candidates by measuring the edit distance between the tag text and a language vocabulary and selecting only the words with a minimum edit distance. In addition, depending on the length of a tag, we attempted to break the tag into multiple vocabulary words. Each candidate generated by the spell checking and tag splitting processes was scored based on how well it matched tokens within the content of the documents that were labeled with this tag. Furthermore, the scores of spelling variations that were used as (i) tags to label documents to which this tag was assigned or as (ii) tags by the users that also made use of this tag were boosted. For all cases where a tag was split into multiple words, we scored the phrase generated by the splitting process using the probability values of English bigrams[3]. Thus, phrases created by the split of the tag text into random words that do no 'go together' were scored lower than valid English phrases. Once this process was complete, the highest scoring variation of the tag text was used for further processing.

Evaluation

For our dataset of 7,754 English tags, the tokenization and spell correction procedure altered 25.03% of the tags. Its accuracy is 97.16% when evaluated on a randomly selected set with 10% of the folksonomic tags. The main source of errors stems from the "richness" of the English vocabulary as derived from the English Wikipedia articles. The unigram language model derived from this document collection includes, as single words, concepts that could be tokenized into multiple words, e.g., *googlemaps, blogpost, macosx, screenprinting, todo, searchengine*, etc. Because the untokenized version of the tag is found in the dictionary, no changes are attempted by the system.

3.3.1.3. Capitalization restoration

In order to restore the proper capitalization of tags that may denote proper names, we compared each tag text with the content of the documents that were labeled using that tag. We note that the document content includes the correct spelling and capitalization information for a tag. We also extend these comparisons to the titles of the labeled documents. For this processing step, we computed a likely capitalization for the tag based on all the documents labeled with the given tag. If a tag is not found within the content of a document, the system fell back to the capitalization computed across all documents. We note that any competing values for a tag's

[3] We used the English Wikipedia articles to create a bigram language model for the English language. We note that trigrams produce more accurate phrases, however, the time and computer memory needed to compute all trigram probabilities seen in this large dataset were exceeding the time and the capabilities of the computers allotted to this task.

capitalization are scored based on the position of the candidate within a document (English headlines capitalize the initial letter of all their content words; Sentences begin with a capitalized word regardless of the correct capitalization of the word). For an accurate understanding of the folksonomic tags, the capitalization of a tag plays an important role during the process of identifying whether the tag is a named entity as well as the class to which the tag belongs to. This is true for tags taken as a whole as well as for words or phrases that are part of a tag.

Examples of tags modified by this processing step include: *linux* → *Linux*, *xhtml* → *XHMTL*, *bbq* → *BBQ*, *javascript* → *JavaScript*, diy → *DIY*, *christian_fiction* → *Christian fiction*, *amish* → *Amish*, *twitter,* → *Twitter*, *bradley/colin* → *Bradley / Colin*, *latex* → *LaTeX*, etc.

Evaluation

The accuracy of this processing step as described above is 89.00% when compared to the manual annotations for a subset of 846 folksonomic tags. We note that errors made by previous tag understanding steps will propagate. Most errors occur because certain tag constituents appear only in document headlines/title and cannot be correctly disambiguated. An additional processing step that may improve these results will identify the correct capitalization of a tag or tag constituent within a much larger set of documents, not only documents labeled with that tag.

All these processing steps transformed the folksonomy from an unstructured set of tags into a collection of phrases, which are correctly spelled, capitalized, and tokenized. Links to the original tags exist. Some of the tags remained unchanged during this process. However, most tags were lexically normalized into well-formed phrases, which were accurately processed by Lymba's suite of NLP tools.

3.3.2. Syntactic understanding of tags

All English tag phrases (as resulted from the lexical understanding step) were processed using the Lymba's part-of-speech tagger, sentence boundary detector, and syntactic parsers (a chunk parse followed by a full syntactic parser).

For the part-of-speech tagging step, preference was given to the NOUN part-of-speech for single word tags, which cannot be tagged within a context. Ambiguities were also resolved by selecting the part-of-speech of the tag as it was marked within the content of the documents labeled with that tag.

The sentence boundary detection step is part of our processing pipeline. It did not modify its input in an overwhelming majority of cases (very few tags spread across sentences).

The syntactic parsing step processed tags with more than one token. This process identifies the type of the tag phrase (NP = noun phrase, VP = verb phrase), its syntactic head as well as any syntactic dependencies between the tag's constituents. This information was later used by Lymba's semantic parser as well as the ontology generation procedure.

We list below several non-trivial examples:

ushistory → *US history* → *US*/NNP *history*/NN → (NP (NNP *US*) (NN *history*))

10.000+words → *10.000 words* → *10.000*/CD *words*/NNS → (NP (CD *10.000*) (NNS *words*))

christopher_hitchens → *Christopher Hitchens* → *Christopher*/NNP *Hitchens*/NNP → (NP (NNP *Christopher*) (NNP *Hitchens*))

toread → *to read* → *to*/TO *read*/VB → (VP (TO *to*) (VB *read*))

Evaluation

For the part-of-speech tagging task, the system's accuracy is 93.26% when the automatically generated output is manually verified for 10% of the folksonomic tags. Most errors are sourced by bad capitalization errors: adjectives whose first letter is capitalized by the previous processing step are wrongly identified as proper nouns (e.g., *international*, *urban*).

Bad part-of-speech tags lead to a bad syntactic parse of the tag. Thus, the system's accuracy for the syntactic parsing of folksonomic tags is 93.02%. We note that the syntactic structure of the tags is not complex, easing the parser's task.

3.3.3. Semantic understanding of tags

The semantic understanding of tags stage covers the understanding of abbreviations and acronyms, the sense disambiguation of tags and the discovery of semantic relations within multi-word tags. The first two processing steps are the most challenging ones, as they require a broad context for the tag usage. Within folksonomies, social tagging systems, it can be argued that tags are primarily used to help the particular end-user who is submitting them (a tag is a set of words that defines a relationship between the online resource and a concept in a user's mind, freely chosen by the user without any formal guidelines). Thus, every user-selected word actually has a unique meaning. However, the increasing popularity of tagging systems and its social, collaborative effort to label existing content enabled users to browse and search vast bookmark collections, which lead to a natural convergence of tags (and their meaning) with few single-use tags (10-15%). Consequently, we depended on the content of the bookmarked documents to provide the context much needed for the disambiguation of each tag. For tags associated with non-textual documents (images, videos, audio files), we used co-occurring tags existent as part of the social bookmarking data.

3.3.3.1. Acronym and abbreviation understanding

The first step in the semantic understanding of folksonomic tags was to disambiguate abbreviations that either form a complete tag or are part of a larger tag. For the purpose of identifying abbreviations, we used Lymba's compiled abbreviations dictionary which comprises of 118,055 distinct abbreviations (almost 25% of the stored abbreviations can be expanded to more than one definition – most ambiguous abbreviation is *SS* with 192 possible definitions within 66 domains). Given the various possibilities for defining, thus disambiguating, an abbreviation, we relied on the tagged document content to determine the correct domain of the

abbreviation. For this purpose, we attempted to link important document concepts to domain descriptions using lexical chains built using WordNet's synset-synset relations. Short chains always indicate strong semantic similarities between the connected concepts, and, thus, the document's subject belongs to a particular topic. Using this information, we narrowed down the set of possible interpretations of the tag. Further disambiguation was done using co-occurring tags and their meanings. Also, by aligning the abbreviation text with the document content (more specifically, its list of simple noun phrases), new definitions for abbreviations were accurately identified and associated with the tags.

For instance, in our dataset collected from the www.delicious.com website, tag *PR* is used to label 1409 documents. In our dictionary, there are 87 distinct definitions for this abbreviation, including, *Press Release, Public Relations, Puerto Rico, Page Rank, Public Radio, Permanent Resident/Residency,* etc. The contents of the documents labeled with this tag were vital to the semantic understanding of the abbreviation. For instance, when used to tag http://prsarahevans.com/2009/06/do-you-have-a-strategy-for-online-comments, *PR* denotes *public relations*, a phrase that appears in the content of the document six times. Other tags used to label the same document in our dataset include *public* and *relations*. On the other hand, when used to label http://www.bbc.co.uk/pressoffice/pressreleases/category/new_media_index.shtml, *PR* refers to *press releases* – also a frequent phrase in the document's content. A less frequent interpretation of *PR* is derived when it is used to tag http://escape.topuertorico.com. We note that none of the three documents included the abbreviation *PR*.

Evaluation

The system's accuracy for this processing step is 95.05% for a randomly selected set of 10% folksonomic tags. We note that most tags are not abbreviations nor do they contain abbreviations or acronyms. Within the entire dataset, the system modified 3.86% of the folksonomic tags by expanding them or some of their components to the definition it considered appropriate. Most of the errors made at this processing step are due to well-established computer concepts such as *HTML*, *ASCII*, *USB*, *PDA*, which are not defined within the contents of the documents they label despite the fact that they may appear within the document. Many of these concepts are defined in many English dictionaries, such as WordNet. Other errors stem from the incorrect classification of the documents into the abbreviation domain. Very few abbreviations were not found in our compiled dictionary of acronyms and abbreviations and were not expanded.

3.3.3.2. Word sense disambiguation

The second step in our semantic understanding process continued the disambiguation process with a multi-stage approach to tag sense identification which assigned each tag or tag concept its corresponding WordNet sense number. For this step, we relied on the content of the documents labeled with the tag. The word sense disambiguation process exploits the linguistic context of the analyzed word. Within documents, this includes the words surrounding the input concept. However, for folksonomic tags, the documents that social bookmarking users labeled provide the needed linguistic context. For tags that appear within their corresponding documents, we use the sense numbers derived by Lymba's word sense disambiguation module during the semantic processing of the documents. For instance, tag *sign* used to label http://www.signingsavvy.com

(Signing Savvy: Your Sign Language Resource) occurs in the document content and its linguistic context on sign language, American sign language, fingerspell, etc. pinpoint to its WordNet sense number 9 (a gesture that is part of a sign language). This sense value is also assigned to the tag concept. For tags that do not appear in the content of their associated documents, that label non-English or non-textual documents, we use the set of co-occurring tags to determine the correct sense of the tag (senses for the tag constituents). For example, when tag *sign* is attributed to http://www.nikonet.or.jp/spring/sanae/report/suusiki/suusiki.htm (Japanese document), we use the set of tags used to label this document to disambiguate *sign*. One of these tags is *mark*, concept synonymous with *sign#1*. Part of another *sign* example, where this tag labels an image file (http://img179.imageshack.us/img179/6307/2172685295d8860567cbb.jpg), its co-occurring tags (*graffiti*, *pics*) pinpoint to its second sense (a public display of a (usually written) message). We note that we use WordNet for our sense inventory. For non-WordNet concepts, we use Lymba's named entity recognizer tool to associate named entity classes to tags. These can be derived from the content of the tag's set of corresponding documents, or based on the grammar rules and lexicons that the module uses (no context is needed for certain named entity classes such as *date*, *number*, *money*, etc). For instance, the tag *christopher_hitchens* is used to label URL http://www.salon.com/news/1998/07/13news.html. The content of the document includes two mentions of this tag (in its normalized formed), both marked as *human* named entities during the document's processing through Lymba NLP pipeline. Tags such as *2009MAY* or *1960s* are easily identified as dates.

Evaluation

The accuracy of the word sense disambiguation step on a randomly selected set of 10% folksonomic tags is 82.51%. There are several sources of error: (1) inherent word sense disambiguation errors caused by semantically close WordNet senses, (2) word sense disambiguation errors within document contents that propagate to the tags, (3) limited linguistic context for certain tags – problem alleviated for analyses of a larger dataset.

For the named entity recognition step, our proposed system achieves an accuracy of 85.81% when evaluated on a randomly selected set of 10% of folksonomic tags. Most errors are due to the fact that named entity tags are not recognized as named entities (e.g., *Twitter*, *Apple*, *Java*, or *BBC*) when the tag is analyzed, but also within the content of the documents they appear in. Fewer errors are caused by the labeling of a tag with the wrong named entity information (e.g., *cycling* as _award, *Dewey* as _town, *rubik* as _town). We note that within the entire dataset, 10.04% of the tags were marked as named entities.

3.3.3.3. Semantic parsing

For single-word tags, the word sense disambiguation processing step produces a semantic representation of the tag and the system is now able to use the extracted information to link the tag with other tags as part of the ontology building process.

For multi-word tags, an additional processing step is required: the semantic parsing of the tag – the discovery of semantic relations that connect the tag concepts – thus completing the semantic understanding of the entire tag. For this final step, we used Lymba's semantic parser, Polaris [3].

It identifies 35 semantic relations (examples shown in Table 3) using a combination of semantic rules and machine learning classifiers. The semantic parser relies of the senses assigned to each word as well as on their syntactic/grammatical dependencies (the syntactic parse of the phrase) to derive the correct semantic relation. Examples of semantic relations identified within tags include TEMPORAL(later,read) for tag *readlater*, PROPERTY-ATTRIBUTE-VALUE(primary,source) and ISA(primary source,source) for tag *primary_sources*, and INSTRUMENT(stick,fight) and PURPOSE(fight,stick) for tag *fightstick*.

Table 3. Semantic relations identified by Lymba's semantic parser Polaris

Relation (Memonic)	Definition	Example
AGENT (AGT)	X is the agent for Y; X is prototypically a person.	[XY] [John] [eats] eggs and ham
ANTONYMY (ANT)	X is the opposite of Y; X is not Y	[XY] A person that is [single] is not [married]; [XY] The [light] contrasts with the [dark]
CAUSE (CAU)	X causes Y	[XY] [Drinking] causes [accidents]
ENTAIL (ENT)	X entails Y; If X, then Y	[XY] Where there's [smoke], there is [fire]
INSTRUMENT (INS)	X is an instrument in Y	[YX] John [broke] the window with [a hammer]; [YX] John [played] the Brandenburg Concerto on [the harmonica]
ISA	X is a (kind of) Y	[XY] [John] is a [student]
KINSHIP (KIN)	X is a kinship of Y; X is related to Y by blood or by marriage	[XY] [John]'s [uncle]
LOCATION DIRECTION PATH GOAL (LOC)	X is the location of Y or where Y take place	[YX] There is [a cat] on [the roof]; [YX] The hurricane [passes] through [Galveston]
MAKE-PRODUCE (MAK)	X makes Y	[XY] [GM] manufactures [cars]
MANNER (MNR)	X is the manner in which Y happens	[YX] John [read] [carefully]; [ran] [quickly]; [spoke] [hastily]
PART-WHOLE (PW)	X is a part of Y	[YX] [faculty] [professor]; [XY] [door] of the [car]
POSSESSION (POS)	X is a possession of Y, Y owns/has X	[YX] [John] owns [a Porsche]; [YX] [John] has [4 acres]
PROPERTY TYPE (PRO)	X is a property type of Y	[XY] [The color] of [the car] is blue
ATTRIBUTE VALUE (VAL)	X is a property/attribute/value of Y	[YX] [The car] is [blue]; [YX] [The color] of the car is [blue]
PURPOSE (PRP)	X is the purpose for Y; Y did something because this person wanted X	[YX] John [swims] for [fun]; Mary [works] part-time [to earn some extra money]
SIMILARITY (SIM)	X is similar to Y	[XY] [Harry] resembles [Zelda]
SOURCE (SRC)	X is the origin or previous location of Y	[XY] [Chilean] [Sea Bass]; [YX] [Student] from [Russia]
TEMPORAL (TMP)	X is the time of Y (when Y take place)	[XY] John [woke up] at [noon]
THEME PATIENT	X is the	[YX] John [painted] [his truck];

RESULT (THM)	theme/patient/result/consumed in/from/of Y	[YX] John [baked] [a cake]
TOPIC CONTENT (TPC)	X is the topic/focus of cognitive communication Y	[YX] John [talked] about [politics] with Mary; [YX] John [said] [he likes the other party]

Evaluation

The overall accuracy of this step is 94.50% when measured on a randomly selected set of 10% folksonomic tags. We note that the evaluation was not restricted to multiple word tags. Thus, errors propagated from previous processing steps are accounted for during this evaluation process. Within the set of analyzed tags, 17.6% of the tags were multi-word concepts. Most of the tags marked with an inaccurate semantic relation understanding were missing relations that would complete the tag's meaning. For instance, *cycling_blogs* (normalized to *cycling blogs*) is marked as having an ISA(cycling blogs,blogs) semantic relation without an additional TOPIC(cycling,blogs) relation. Fewer errors were cause by the "abnormality" of the tag. Because the word order is reversed, Polaris cannot derive the correct semantic relations that link the tag concepts (e.g., *things_japanese, Radio_Online*).

This linguistic processing step completed the process of understanding the tag semantics and its representation into a machine-readable format that was further exploited to automatically generate an ontology.

3.4. Deriving the Folksonomy's Structure from Tag Semantics

Once we completed the process of understanding what each tag represents, we shifted our focus to the derivation of the folksonomy structure from the tag semantics. We began by connecting tags using EQUALITY and SYNONYMY relations.

EQUALITY relations were created between tags with the same lemma, part-of-speech, and sense number. These are relations connecting highly correlated tags. Non-trivial examples include: EQUALITY(*activity*, *activities*), EQUALITY(*after-effects*, *AfterEffects*), and EQUALITY(*opinion*, *Opnion*). In addition to the linking identical tags assigned to multiple bookmarks, this relation type links tags that are syntactically normalized to the same form, tags that are tokenized (including capitalization) to the same form and misspelled tags to their correct form tags.

SYNONYMY relations were assigned to pairs of tags that have the same synset id. These tags belong to the same synset in WordNet (for single word tags), thus deemed synonyms within WordNet. The synset id is derived based on the lemma, part-of-speech and sense number of the tag. For instance, tags *Archeology* and *Archaeology* are part of the same WordNet synset (id: 06144081); *OS* and *operating.system* are synonyms within a WordNet synset (id: 06568134). Furthermore, we used the named entity and abbreviation information to identify SYNONYMY relations between tags that refer to the same concept using different wordings. This is extremely useful for non-WordNet concepts. Examples includes SYNONYMY(*LA*, *losangeles*), SYNONYMY (*nyt*, *nytimes*), etc. We also created SYNONYMY relations between multi-word tags that satisfy the following criteria: they have synonymous constituents that are linked by the

same semantic relation. All SYNONYMY relations connect semantically similar tags. These links are not as strong as the EQUALITY relationships.

In addition to EQUALITY and SYNONYMY relations, we implemented automatic procedures that derive additional tag-tag relations.

An initial set of ISA relations was created between all named entity tags and their corresponding WordNet synsets that describe the name of the entity class. For instance, there is an ISA relation between tags *OracleCorporation* and *organization*. Another example includes ISA(*davidfosterwallace*, *person*). We note that most named entity tags are not defined within WordNet and these ISA relations are vital in describing the hierarchical structure of the folksonomy. These relations denote a directional semantic subordination of their arguments.

By mapping our SYNONYMY clusters to WordNet, we were able to add to our ontology existing WordNet relations that link two folksonomic tags. This procedure added 23.66% of the total number of relations to the ontology. Examples include ISA(*vegan*, *vegetarian*), ANTONYMY(*peace*, *war*), PART_WHOLE(*Businesses*, *markets*), ENTAIL(*proofreading*, *+read*), SIMILARITY(*important*, *general*), and DOMAIN(*light*, *physics*).

We also built lexical chains of size two between tags. They are of the form $tag_1 - rel_1 \rightarrow synset - rel_2 \rightarrow tag_2$, where tag_1 and tag_2 are part of our folksonomy, rel_1 and rel_2 can be any two semantic relations and *synset* is part of WordNet. Given these lexical chains, we used Lymba's semantic calculus rules [5], which derive new semantic relations by combining two semantic relationships, to add new tag-tag relations to our folksonomic ontology. 41.02% of the ontological relations were added using this procedure. For instance, ISA(*integration*, *events,*) is a relation derived from the combination of ISA(*integration*, group_action/NN/1) and ISA(group_action/NN/1, *events,*). We note that the concept connecting the two tags is not part of the folksonomy. If *synset* were itself a tag, then the semantic calculus rules would create a redundant relation, which would be removed by further processing. Similarly, PART_WHOLE(*lobby*, *hotels*) is derived from PART_WHOLE(*lobby*, building/NN/1) and ISA(building/NN/1, *hotels*).

For complex tags, we used their semantics to find related tags. For instance, for tags of the form *modifier head* where there is a semantic relation between *modifier* and *head* relation and where *head* constitutes a folksonomic tag, we added an ISA relation between the *modifier head* and *head* tags. The relation linking *modifier* and *head* can be a PROPERTY_ATTRIBUTE, PART_WHOLE and even a TEMPORAL relation. This procedure accounts for 17.12% of the total relations added to the ontology. Examples include ISA(*book-cover*, *covers*), ISA(*theoryofmind*, *theory*), and ISA(*photoshoptutorials*, *tutorials,*).

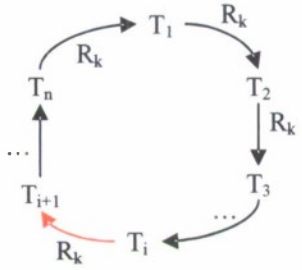
For complex tags of the form $modifier_i$ $head_i$ where there exists a semantic relation REL ($modifier_i$, $head_i$), (i=1,2), we explored any semantic connections between $modifier_1$ and $modifier_2$ as well as between $head_1$ and $head_2$ in order to derive semantic links between $modifier_1$ $head_1$ and $modifier_2$ $head_2$. We implemented to following classification rules for these tags:

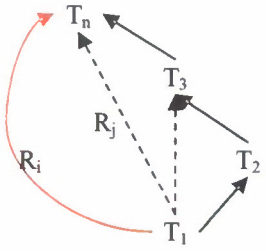
- If ISA(*modifier*$_1$,*modifier*$_2$) and ISA(*head*$_1$,*head*$_2$), then we add a new ISA relation between the tags;
- If ISA(*modifier*$_1$,*modifier*$_2$) and SYNONYMY(*head*$_1$, *head*$_2$), then a new ISA relation is generated between the two tags;
- If SYNONYMY(*modifier*$_1$,*modifier*$_2$) and ISA(*head*$_1$,*head*$_2$), then we add a new ISA relation between the tags;
- If SYNONYMY(*modifier*$_1$, *modifier*$_2$) and SYNONYMY(*head*$_1$,*head*$_2$), then we create a new SYNONYMY relation between the two complex tags;
- If SYNONYMY(*modifier*$_1$,*modifier*$_2$) and REL(*head*$_1$,*head*$_2$), where REL could be any semantic relation, then a new REL relation is added to the ontology.

Examples include ISA(*build-solar-panel*, *create-solar-panel*), SIMILARITY(*socialnetworks*, *socialweb*) (based on the SIMILARITY(*networks*,*web*) which was derived using Lymba's semantic calculus rules – both nouns are derivations of the concept web/VB/1).

Once we derived this rich set of semantic relations between the folksonomic tags, we performed few sanity checks to ensure that our ontology has a consistent structure that can accurately support any applications involving the input folksonomy. These checks include a consistency check that identifies and resolves any conflicts as well as a redundancy check.

For our generated ontology, a conflict is detected when transitive relations form a dependency cycle (left hand side figure below) or when the relation closure procedure derives a relation R_i between two tags already connected by an R_j relation (R_i != R_j) (right-hand-side figure below). The resolution procedure identifies the lowest confidence relation among the relations forming the cycle (in the first case) or between R_i and R_j and removes it from the ontology.





A redundant relation is a tag-tag relation derived by the classification rules described above that can also be derived by the relation closure procedure. Because there are other means of generating these relations, we removed all redundant relations produced in the relation generation step.

The resulting ontology is a rich graph with nodes that represent clusters of synonymous tags and labeled directed links that denote the semantic relations that connect the folksonomic tags. Projections of this graph, which include only relationships such as ISA and PART_WHOLE, reveal hierarchical organizations of the folksonomy.

Evaluation

For our social bookmarking dataset, our system created 9820 EQUALITY clusters for the 8460 folksonomic tags. Most tag strings that belong to multiple EQUALITY clusters are abbreviations expanded to different definitions for different bookmarks (e.g., *ST*, *OS*, or *AI*). However, there are EQUALITY clusters that combine multiple unique tag strings (e.g., *tutorial*, *tutorials*, and *tutorials,*) – these tags were normalized (lexically, syntactically, and semantically) to the same concept.

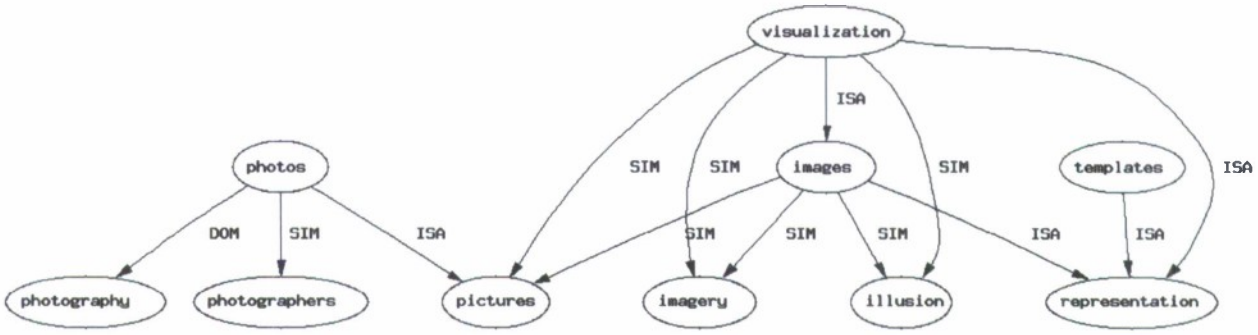
Within the same dataset, our prototype implementation derived 8801 SYNONYMY clusters. Most of the tag strings that find themselves within different EQUALITY clusters belong to different SYNONYMY clusters also. The largest SYNONYMY clusters groups 133 (user,document,tag) triplets where tag can be *car*, *automobiles*, *auto*, *autos*, *cars*, or *automobile*. Other large SYNONYMY clusters include {*movies*, *movie*, *Movies:*, *film*, *films*}, {*gadgets*, *widget*, *widgets*, *gadget*, *appliance*}. The SYNOYMY clusters are determined by the semantic understanding of each tag (associated with a certain bookmark). Thus, any errors made by the system when creating the clusters of synonymous tags were caused by mistakes made during earlier processing stages, most notably the word sense disambiguation step.

Among these SYNONYMY clusters, our system identified 5439 ontological relations using the classification procedures described above. This set of relations uses 11 types of semantic relations. The most frequent is ISA with a total of 3869 instances, followed by SIMILARITY (600 instances), PART_WHOLE (429 links) and others such as DERIVATION, DOMAIN, ANTONYMY, etc. The SYNONYMY cluster that is linked the highest is {*humans*, *person*, *human*} which participates in 89 semantic relations. The most prolific source of semantic relations is WordNet when combined with Lymba's semantic calculus rules. There were 1778 ontological relations derived using this procedure.

3.4.1. Folksonomy Visualization and Browsing

Given the folksonomy's semantic structure made explicit by the ontology we induced from the advanced processing of tags and well as automatic derivation of relations between tags, we began to build a tool that allows users to search, browse and visualize the derived ontology.

Our initial visualization prototype displays the complete ontological graph: all semantic relations are displayed as directed links between tag nodes, each denoting a SYNONYMY tag cluster. We used as node labels single tags randomly selected from within the node's corresponding semantic cluster. We note that the folksonomy structure need not be a connected graph (i.e. there may exist two folksonomic tags, which cannot be linked by chains formed with the ontological relations). Below, we show one of the folksonomy structure's connected components. We used the Graphviz software program to obtain this figure (http://www.graphviz.org).



A more sophisticated visualization, browse, and search tool is currently under development. This tool will highlight the hierarchical structure of folksonomy using the ISA and PART_WHOLE relations identified by our prototype system.

Each ontological node will be represented by the collection of tags forming the SYNONYMY cluster, all accompanied by the number of hierarchical semantic relations in which they

participate. In addition to the tree-like structure, we shall display tag details, including all relations in which the tag participates, the social bookmarking data surrounding the tag and the semantic understanding automatically derived by our system. A tag search box will also be added to the tool. Furthermore, we shall enable an editing feature, which will allow users to modify the automatically generated ontology, by adding or deleting tag-tag relations and by modifying tag characteristics. Above, we show a portion of the generated ontology using our initial version of the tool.

References

[1] Adrian Novischi, Munirathnam Srikanth and Andrew Bennett. 2007. *LCC-WSD: System Description for English Coarse Grained All Words Task at SemEval 2007*. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pages 223–226, Prague, Czech Republic.

[2] Adrian Novischi, Dan Moldovan, Paul Parker, Adriana Badulescu, and Bob Hauser. 2004. *LCC's WSD Systems for Senseval 3*. In Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, July 2004.

[3] Adriana Badulescu and Munirathnam Srikanth. 2007. LCC-SRN: LCC's SRN System for SemEval 2007 Task 4. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, June 23-24, 2007.

[4] Golder, S., and Huberman, B.A.: *The Structure of Collaborative Tagging Systems.* HP Labs technical report. (available in http://www.hpl.hp.com/research/idl/papers/tags) 2005

[5] Marta Tatu and Dan Moldovan. 2006. *A Logic-based Semantic Approach to Recognizing Textual Entailment*. In Proceedings of COLING/ACL 2006, Sydney, Australia, July 2006