Semantic Relation Extraction from Socially-Generated Tags: A Methodology for Metadata Generation

Miao Chen
Syracuse University, USA
mchen14@syr.edu

Xiaozhong Liu
Syracuse University, USA
xliu12@syr.edu

Jian Qin
Syracuse University, USA
jqin@syr.edu

Abstract

The growing predominance of social semantics in the form of tagging presents the metadata community with both opportunities and challenges as for leveraging this new form of information content representation and for retrieval. One key challenge is the absence of contextual information associated with these tags. This paper presents an experiment working with Flickr tags as an example of utilizing social semantics sources for enriching subject metadata. The procedure included four steps: 1) Collecting a sample of Flickr tags, 2) Calculating co-occurrences between tags through mutual information, 3) Tracing contextual information of tag pairs via Google search results, 4) Applying natural language processing and machine learning techniques to extract semantic relations between tags. The experiment helped us to build a context sentence collection from the Google search results, which was then processed by natural language processing and machine learning algorithms. This new approach achieved a reasonably good rate of accuracy in assigning semantic relations to tag pairs. This paper also explores the implications of this approach for using social semantics to enrich subject metadata.

Keywords: relation extraction; tags; search engine; social semantics; metadata

1. Introduction

The recent social tagging movement has generated abundant semantic resources for representing the content of information objects. Unlike traditional subject indexing performed by trained librarians, the socially-generated semantic tags are created by users who want to assign tags to the information objects of their interest. While these tags are sometimes erroneous and ill-constructed (Guy & Tonkin, 2006; Mathes, 2004; Michlmayr, 2002) this newfound wealth of social semantics has become a mining ground for discovering and understanding social networks and cultural taste (Liu et al., 2006; Mika, 2005), ontological structures (Schmitz, 2006), and various semantic relationships among the tags (Rattenbury et al., 2007).

Subject representation as one important area in metadata description may employ social semantics or controlled semantics. The two types of semantics can benefit each other in a profound way as Qin has discussed (2008). On the one hand, social semantics as empirical knowledge can contribute to controlled semantics through testing it and thus learning from it. On the other hand, social semantics provides a valuable source of empirically-derived knowledge to enrich and validate controlled semantics (Qin, 2008). We are facing, however, a number of challenges in accomplishing these goals. One such challenge is the methodology.

Tag mining methodology includes a wide variety of techniques and algorithms used to acquire, preprocess, parse, and analyze tag data. Before tag data becomes usable for mining tasks, it needs a series of linguistic, syntactic, and semantic processing. This processing is often computationally intensive and requires linguistic and semantic sources to be adapted to the mining techniques and tasks. Research on mining social tags to discover semantic patterns and relationships has applied machine learning, clustering, natural language processing, and other techniques (all which are reviewed in the next section).

A major weakness (among other flaws) of user-generated tags is the lack of semantic relations between terms, which are often represented in controlled semantics as broader, narrower, and related terms; or, in ontologies as relations between classes of concepts and instances. While it is

impractical to expect users to categorize tags or provide semantic relations in the same way as librarians do for controlled semantics, it is possible to extract semantic relations using computational methodologies. The study reported in this paper is an attempt to address this methodology challenge. By using Flickr's tags as the source, we applied natural language processing (NLP) and machine learning techniques, in addition to Google search results, to the processing and analysis of Flickr tag data. The goal of this research has been twofold: 1) to experiment with an approach employing NLP and machine learning techniques combined with Web search results to provide the context of tags for extracting semantic relations from social semantics; and 2) to evaluate the effectiveness of this methodology. The long-term goal has been to develop effective methods for meshing up social and controlled semantics that can be used for subject metadata representation of digital objects and resources.

2. Literature Review

Semantic relations between concepts or entities exist in textual documents, keywords or key phrases, and tags generated in social tagging systems. Relation extraction refers to the identification and assignment of relations between concepts or entities. Automatic extraction of semantic relations has a wide range of applications in knowledge organization and information retrieval. Relation extraction can explore relationships that are implicit to underlying data and then add new knowledge to the different domains.

Previous studies have focused on relation extraction between entities from (document) textual resources. In traditional relation extraction, the sources of entities usually come from terms in unstructured documents such as Web pages or structured documents such as relational databases. A wide variety of data sources have been used in relation extraction research, e.g., Web pages (Brin, 1998), corpus (Bunescu & Mooney, 2007), and socially generated Wikipedia articles (Nguyen et al., 2007). The semantic and linguistic sources for exploring relations can be a corpus containing the context of entities, and this context information can serve as the basis of relation assignment.

No matter which data sources are utilized in relation extraction, it is necessary to meet three requirements: 1) a collection of data (entity) sources from which semantic relations will be extracted, 2) a semantic or linguistic source in which the context for relations is provided, and 3) algorithms for automatic execution of processing operations. How well a relation extractor performs is determined mainly by the context sources and algorithms. Context containing entities or concepts play a critical role in ensuring the precision of text relation extraction since this provides the source in which covert relations may inhabit.

While text relation extraction relies heavily on the context, current research on tag relation extraction rarely includes context information in the procedure. Tag relations are extracted by applying statistical methods to derive relations from tag co-occurrences, similarity computations, and usage distribution. Examples of these types of studies include a hierarchical taxonomy built from the Deli.cio.us and CiteULike tags by using cosine similarity of tag vectors (Heymann & Garcia-Molina, 2006), and an ontology generated from Flickr tags using statistical methods (Schmitz, 2006) that in turn was based on Sanderson and Crofts' (1999) model for the co-occurrences of tags. For each frequently co-occurring pair of tags, the model was applied to determine whether or not there was a hierarchical relation between them. Subsequently, a hierarchical structure of tags became an ontology. Rattenbury et al. (2007) presented an approach of identifying event and place tags from Flickr. The assignment of tags' semantic types was learned from patterns of temporal and spatial tag usages employing statistical methods. When contrasted with the three requirements of text relation extraction, it becomes apparent that the second requirement for context is missing from these tag relation extraction experiments.

Although the abovementioned methods have achieved varying levels of success, the absence of context information in these methods limits not only the accuracy of processing but also the scalability of automatic relation extraction. Our strategy in addressing this limitation was to add

context to tags. By tracing tags to the context where they might have originally appeared or commonly been used, we could explore the context that would assist us to extract accurate and reliable relations. The methodology we employed involved using external document resources that have sentences containing the tags in the source data. Relations were then extracted from these documents and assigned to related tags.

Extracting semantic relations from documents is not a new area of research; in fact, a large number of studies on extracting relations from text (including Web pages) and corpus have been published in the last two decades. Relation extraction generally involves two primary parts: 1) the natural language processing (NLP) part, and 2) the machine learning part. NLP techniques are applied in order to identify entities and relation indicators from texts. Machine learning algorithms are implemented to learn features of relations, and assign relations to entities whose relations are not yet known. Text relation extraction also involves entity extraction for identifying entities or concepts (Brin, 1998; Iria & Ciravegna, 2005; Nguyen et al., 2007; Roth & Yih, 2002).

The NLP part of relation extraction is a process by which the text processing may be performed at different levels throughout different stages using either shallow processing or deep processing. Shallow processing involves sentence segmentation, tokenization, part of speech (POS) tagging, and chunking of the text being processed—which is used to identify phrases and chunks (Bunescu & Mooney, 2007). An example is the study by Roth and Yih (2002), where shallow parsing was used to segment sentences and to identify entities and relations. Deep processing builds a parsing (or dependency) tree by identifying the shortest-path dependency of language components in sentences. This NLP technique is useful when the context of pairs of entities needs to be processed. In such cases, the words located before, between, and after these entities are used directly as vectors for matching patterns of relations (Agichtein & Gravano, 2000). The question of whether to use a shallow or deep level of text processing is determined by the design of experiment(s) and algorithm of machine learning. If shallow processing is sufficient, then there is no need to use deep processing (Zelenko et al., 2003).

Machine learning performs a different role in relation extraction. As computer algorithms, machine learning is dependent upon features (variables) representing objects as the input into learning models. The features needed for machine learning may be entity types, words, phrases, part of speech, chunks, tags, etc. from the context sentence or sentence part (Bunescu & Mooney, 2007; Culotta & Sorensen, 2004). From samples (context containing pairs of entities) whose features and relation types are already known, machine learning generates patterns of different relations based on features. Subsequently, the generated patterns can be applied to new contexts with unknown relations and derive meaningful relations. Commonly used machine learning models include the support vector machine (SVM) (Bunescu & Mooney, 2007; Culotta & Sorensen, 2004; Zelenko et al., 2003), clustering (Agichtein & Gravano, 2000), undirected graphical models (Culotta et al, 2006), and decision tree (Nahm & Mooney, 2000).

A review of previous studies shows that past research in tag relation extraction has rarely used contextual sources for relation recognition and has seldom utilized techniques from text relation extraction. Tag relation extraction as a special case of relation extraction does not need entity extraction (because tags are not sentence-based documents) as does regular text relation extraction. To leverage the social semantics power for subject metadata description, we are faced with challenges brought about by the lack of context information in tag sources. Solving this problem is a critical first step to successfully deploying social semantics in subject metadata description. We will introduce the details of the proposed methodology for improving tag relation extraction in Section 3, the experiment using our methodology in Section 4, the results and performance in Section 5, and discussion of the results and conclusions in Section 6.

3. Methodology

In this section, we introduce our methodology in detail and explain the process of extracting relations between Flickr tags. Two sources are critical in this process: the source of entities and

the source of context. Since entities have already been "extracted" by taggers, we instead focus on obtaining the context of the tags in our sample data by using search results from a general search engine.

As mentioned above, a major challenge in extracting tag relations is the lack of context information for the tags, which makes them insufficient and difficult for the relation extraction task. An example is a photo in Flickr that has been assigned four tags: *Shangrila* (a remote area in southwest China), *Mountain*, *Yunnan* (one of the provinces in China), and *River*, as shown in FIG. 1. Since the context is the photo itself and separated from the tags in the search system (i.e., image-based search is still not available in most search systems), the four tags could have a wide variety of contexts for interpretation when separated from the photo they describe.

FIG. 1. A photo in Flickr with four tags: Shangrila, Mountain, Yunnan, River.

From the perspective of relation extraction, photos do not provide sufficient context for tags and the relations between tags are not explicit. Due to technological limitations, it is difficult to process images in order to acquire semantics. Compounded by the technology limitation is the tagging practice that does not label any relations between the tags, e.g., relation "Shangrila is located in Yunnan" is information separate from either the photo or the tags. Acquiring tag relations without context information is analogous to a simple keyword search on the Web—the precision and recall can be very problematic. These predicaments led us to seek external text resources such as search engine results as a solution to obtaining the context of tags. A unique advantage of using tags to extract relations is that the entities are already "extracted" by human taggers and so the final error rate can be reduced by avoiding the errors that are propagated by the entity extraction process.

3.1. Identification of Problem

Given a set of tags from social tagging Web sites, our task was to discover relations between any two tags that frequently co-occurred. We defined our tag set as (Tag_1, Tag_2, Tag_3, Tag_4, ..., Tag_n) ($n \in N$) and used statistical techniques to identify frequently co-occurring pairs of tags in the tag set. The selected tag pairs were then deposited in a new set called "tag pairs." A tag pair may be represented as *pair* (Tag_x, Tag_y), where Tag_x and Tag_y meet the requirement that both tags frequently occur together. Once the set of tag pairs was constructed, the next step was to identify the relation between Tag_x and Tag_y for each pair in the set.

To precisely and effectively identify relations between pairs of tags, the critical component is the context of tag occurrence. We determined that an effective method was to put the tag pairs back into context by employing results from a general search engine, and then applied natural language processing and machine learning techniques to extract relations from that context. The task at this stage included finding the context for tag pairs and building a classifier for relation assignment. For a tag pair (Tag_x, Tag_y), the relation was defined as R_{xy}, representing a single type of relation between Tag_x and Tag_y.

3.2. Assumptions

We made two assumptions regarding to tag relations. First; if two tags frequently co-occurred, there ought to be some type of relation between them or else they would not be frequently tagged together by users. A high frequency of co-occurrences is not coincidental; rather, it underscores the possibility of some connection between the tags. For example, since "San Francisco" co-occurred frequently with "bay area," we assumed that there was a strong possibility that a relation existed between the two tags. From our knowledge, the two have a relation that San Francisco "*is located in*" the bay area.

The second assumption: there is only one single relation between the two tags in a pair. It is possible that the two have more than one relation, e.g., San Francisco can be "*located in*" the bay area (San Francisco Bay) or San Francisco can be "*located in the north part of*" the bay area. When our human coders were assigning relations to tag pairs for the training data set, they assigned the most general and higher-level relations to the tag pairs. Using the San Francisco example, the relation is "*located in*" (since it is a higher level relation) that includes the instance of "*in the north part of.*" This assumption facilitated the extraction of more features for the learning model.

3.3. Selection of Tag Pairs

We downloaded 28,737 photos with 289,216 accompanying tags about landscape from Flickr– which contained 21,443 unique tags. These all co-occurred with (and are about) the tag "landscape." We used an index of mutual information to find pairs of tags that frequently co-occurred. The mutual information (MI) index between any two tags was calculated based on the co-occurrence between two tags, which was also used to describe and normalize the co-occurrences between two tags. The MI index represents the degree of relatedness in candidate tag pairs (Tag_x, Tag_y), i.e., the higher the MI scores, the more closely related the two tags are. The MI index was calculated by using the well-established formula below (Shannon, 1948):

$$MI(Tag_X, Tag_Y) = P(Tag_X, Tag_Y) \cdot \log \frac{P(Tag_X, Tag_Y)}{P(Tag_X) \cdot P(Tag_Y)} \qquad \text{[EQ. 1]}$$

For the tag set (Tag_1, Tag_2, Tag_3, Tag_4, …, Tag_n), we calculated the MI scores for any two tags, which resulted in an n x n matrix. Tag pairs with low MI scores were removed from the matrix and the remaining high MI score pairs were retained.

3.4. Relation Extraction

Having prepared tag pairs for relation extraction, the next step was to identify the context for tags and generate machine learning models for relation extraction. This process involved 1) entering a tag pair in a search engine query, 2) obtain search results, and then process the search results with NLP tools, 3) establish learning relation patterns from samples with known relation types, and then 4) derive candidate relations for tag pairs. FIG. 2 demonstrates the process of tag relation extraction.

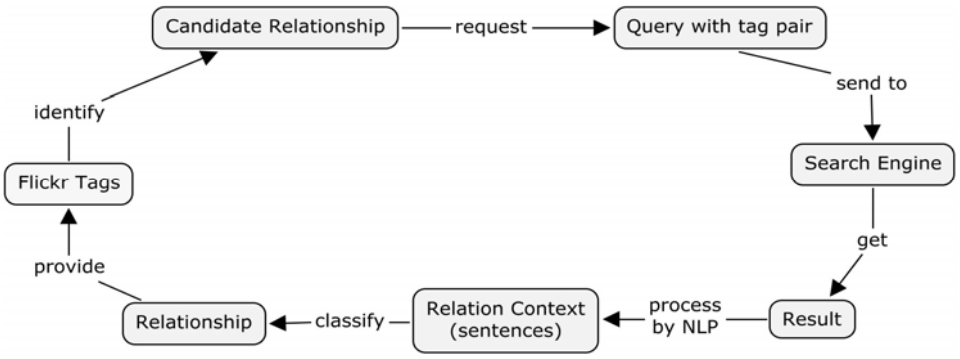

FIG. 2. Tag relation extraction process.

The tag pair query in the general search engine returned a list of results with a title and brief description for each resource in the result set. We assumed that such search results would provide the context for the tags if both tags in a pair appeared together in the results that were highly relevant to the query. If sentences from the search results contained both Tag_x and Tag_y, the sentences were then considered as the context of relation R_{xy} between Tag_x and Tag_y. Although not every returned sentence contained both Tag_x and Tag_y, the only ones needed contained both tags to use as context. Sentences meeting this criterion were selected for the context sentence collection.

Sentences in the context sentence collection were then parsed and chunked using NLP techniques. We applied the deep processing technique because it enabled us to learn more about the features of the context. The NLP processing returned a parsed sentence with part-of-speech tags of words and chunking tags of phrases. For example, a sentence "The largest city in the Sonoran Dessert is Phoenix, Arizona" is parsed into a tree (shown in FIG. 3).

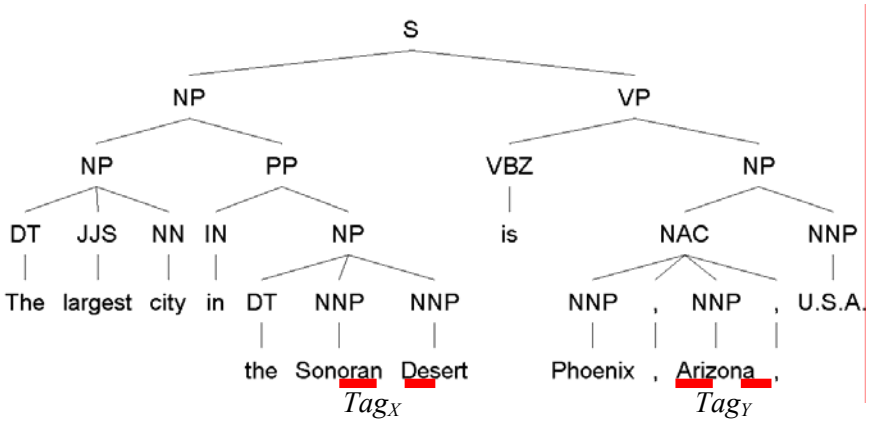

FIG. 3. Parsed tree of a sentence.

As already mentioned, a list of statistical and semantic features can be extracted from natural language processing results. Following Bunescu and Mooney (2007), we used the features of context before Tag_X, between Tag_X & Tag_Y, as well as after Tag_Y. The types of features we chose included: word (the word was processed by a Porter stemming algorithm for stemming), part of speech (e.g. verb, noun, and preposition), chunking, dependency subtree, and the distance between source feature and target feature. In the example sentence from FIG. 3, Tag_X is Sonoran Desert, Tag_Y is Arizona, and the goal is to find the relation between the two tags. The features scrutinized included: (Verb, between_Tag_X_Tag_Y), (verb, is) (DT, before_Tag_X, distance-1), (Tag_X, exist_in_NP), (Tag_Y, exist_inVP), (Tag_X, Tag_Y, lowest_common_father_S), and so forth.

Verb, between_Tag_X_Tag_Y) means that the verb between Tag_X and Tag_Y was taken as one feature, and other listed features can be similarly interpreted.

Once the relation between sample tag pairs was known, features of the tag context were input into machine learning algorithms to generate patterns of different relations. The machine learning algorithm applied the decision tree technique and features were selected to build a classifier for relation extraction. When a new tag pair was identified, the processing went through the above steps for identifying the context through search engine results and natural language processing. The resultant features were then entered into the classifier which later returned the relation type for the tag pair.

4. Experiment

As described in Section 3.3, the dataset contained 289,216 tags. The criterion for including a tag in the dataset was that if a tag appeared together with "landscape" for one or multiple photos, this tag would be included in the tag set. This selection process yielded 21,443 unique tags in the landscape domain.

Each tag pair in the tag set was then computed to generate a matrix of mutual information scores. Tags appearing less than 5 times in the tag set were deleted in order to reduce computation cost. After ranking the mutual information scores (from high to low) in tag pairs, the first 3,000 tag pairs were selected to form the tag pair set. Some example pairs are shown in the following table:

TABLE 1. Examples of tag pair's mutual information.

Tag_X	Tag_Y	*Mutual Info*
bay area	golden gate bridge	0.071521409
Backpacker magazine	CDT PROJECT	0.05926981
beach	ocean	0.058470874
beach	Florida	0.01961479
Beach Houses	vacation	0.015982523
aguila	snake	0.011943919
Acadia	Acadia National Park	0.011012993

As with the examples above, if Tag_X and Tag_Y have a high mutual information score, we can assume that there exists a strong relationship between Tag_X and Tag_Y, then marked as "Tag_X, candidate_relationship_?, Tag_Y." We used the following algorithm to identify candidate relationships:

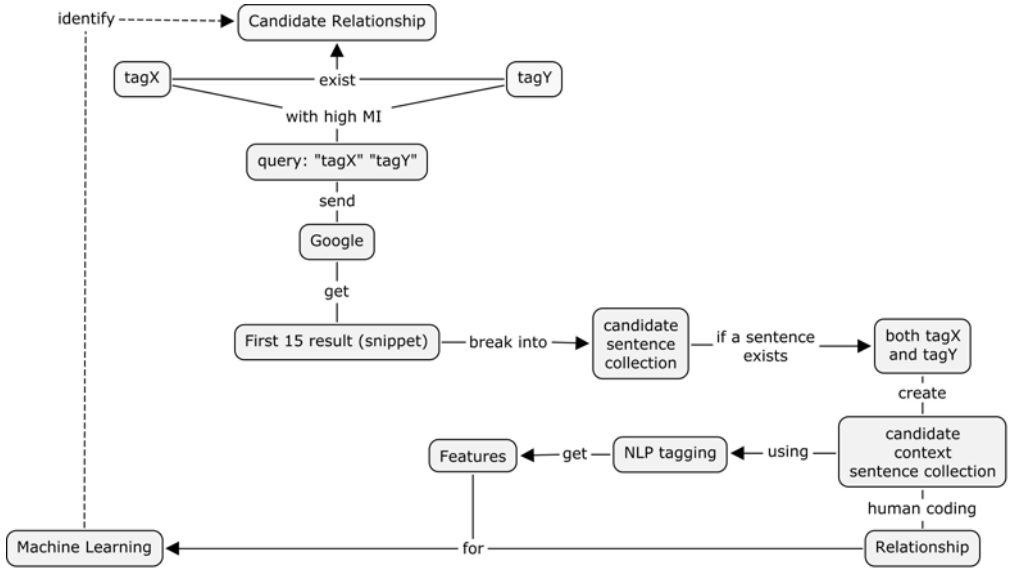

FIG. 4. Process of identifying candidate relationships between tags.

We chose Google as the general search engine whose results would provide context for relation extraction. Google has an API that provides snippet of retrieved sentences. By sending a tag pair (Tag_X, Tag_Y) as a query ("Tag_X" "Tag_Y") to the API, we received a snippet of each result. We used quotation marks in the query because both Tag_X and Tag_Y might be phrases rather than single words, and quotation marks ensure a more exact match for target phrases.

Snippets of the first 15 search results for tag pairs were exported and processed by a sentence boundary tool to identify sentences and put them into candidate sentence collection. The program tested each sentence to see whether or not it contained both Tag_X and Tag_Y. If so, this sentence would be included in the candidate context sentence collection. There was often more than one sentence in the snippet satisfying this requirement for contextual information.

Two human coders manually marked relations for a small portion of the tag pairs in the sentence context collection. Since one tag pair might have more than one context sentence—while only one relation type can be assigned to a tag pair regardless how many context sentences it might have—the most general and high-level relation was assigned to the tag pair. The manual coding produced eight types of relations: 1) *is-a-measure-of*, 2) *is-located-in*, 3) *induces*, 4) *is-induced-by*, 5) *is-style-of*, 6) *is-of*, 7) *is-for*, and 8) *is-a-method-of*. Examples of the relation between tag pairs and context sentences are presented in the following table:

TABLE 2. Human coded relations and context sentences.

tagX	**relation**	***tagY***	**Context Sentence**
2-deoxy-d-glucose	induces	effect	Effect of 2-deoxy-D-glucose on cell fusion induced by Newcastle disease and herpes simplex viruses.
2-DG	induces	effect	Effect of peripheral 2-DG on opioid and neuropeptide Y gene expression.
Action	is-induced-by	anticonvulsant	Pharmacokinetic modeling of the anticonvulsant action of phenobarbital in rats. J Dingemanse, JB van Bree and M Danhof.

tagX	**relation**	***tagY***	**Context Sentence**
Action	is-induced-by	epilepsy	From Epilepsy Action, the UK's leading epilepsy charity.
Acadia	is-located-in	maine	Use this vacation and travel guide to the Downeast and Acadia region of Maine to plan your vacation, business trip or just for fun.
Alabama	is-located-in	America	The Alabama Location Map indicates the exact geographical position of the states of the United States of America.
America	is-located-in	san francisco bay area	Boy Scouts of America, San Francisco Bay Area Council • 1001 Davis Street, San Leandro, CA 94577-1514, (510) 577-9000.

In the preliminary experiment, we chose three relation classes (for 121 cases) for machine learning tasks from human coded relations: "*induces,*" "*is-induced-by,*" and "*is-located-in.*" Among the 121 sentences, part of them were used as the training set for feature extraction and model building, and the remainder were used for evaluation.

We applied the Stanford parser for parsing and chunking in the NLP phase. This step was to prepare for the machine learning part. The parsing of context sentences generated candidate features for machine learning, and when combined with features and relation labels we were able to then conduct training to derive a classifier for relations. A decision tree was the algorithm for selecting features and generating patterns for different types of relations. The resultant classifier was then ready for accepting new context for tag pairs and outputting relations.

Finally, we examined the methodology by sending new tag pairs to the trained model. We withheld the other context sentences as a testing set, and input the sentences to the NLP processor and classifier accordingly. The classifier returned the relation of each tag pair as an automatic relation extraction result. Since we had the human coded results, we compared them with the machine learning results and evaluated the performance.

5. Results and Analysis

Our preliminary experiment extracted 2401 unique features from 121 context sentences. We used a ten-fold cross-validation to evaluate the result (Table 3). The evaluation result of our method displayed in Table 3 shows an 83.72% rate of correct classification / tag relation instances. While the sampling size of tag data and the number of human coded relations could not be as large as we would have liked, this approach appears to be a promising methodology. The introduction of external sources allows for objectively identifying contextual information for context-less tag data and thereby improving the accuracy and reliability of relation extraction.

TABLE 3. Evaluation result of the preliminary experiment.

	is-located-in	*is-induced-by*	*induces*
is-located-in	90	2	1
is-induced-by	10	12	3
induces	1	4	6
Correctly Classified Instances 108		83.72 %	
Incorrectly Classified Instances 21		16.28 %	

This result also suggests that using external sources for context information can help detect data anomalies in the tag pairs that have a high MI score. We discovered from our experiment that a high MI score did not necessarily mean that Tag_X and Tag_Y always had direct semantic

relations. Some tag pairs did not appear in any sentence in Google search results and no context was found containing the tags. For example, the two tags Tag_X = "all rights reserved" and Tag_Y = "Canon EOS 350" had a high MI score, but neither of these two tags appeared together in Google search results; this suggests that no context sentence existed for the two tags. We are unsure at present how the two tags might be related, but it is possible that they are indirectly related. If, for example, they are both related to a third tag in a meaningful way, then they could be related to each other statistically but not semantically. Consequently, the two tags were semantically unrelated and the pair was removed from the tag pair collection to ensure the meaningfulness of tags and their relations.

We also discovered that NLP algorithms can provide flexible and powerful features for relation identification. For instance, a syntax level feature can be helpful for identifying the "*is-located-in*" class in an example pattern such as Tag_X, Tag_Y, Zip Code or Tag_X prep Tag_Y (*prep* could be "in," "with," or "by"), where Tag_X could be a city name and Tag_Y a state name. The NLP algorithms then can be expanded and explored with more semantic feature types and other machine learning algorithms.

6. Conclusion

Tags are a special type of subject metadata as well as a rich, powerful vocabulary source. Extracting relations between tags is the first step toward automatic subject metadata creation. An important contribution of this study was the introduction of external resources as a solution to the problem of context-less tag data. Through combining NLP and machine learning techniques we developed a set of algorithms and procedures for automatically processing the external resources, using the output to provide more objective, reliable context information for tag relation extraction.

The methodology developed in this study can be applied to larger-scale research in the future as well as in research fields beyond tag relation extraction. For example, the processing and categorization of unstructured text can benefit from this methodology, as can automatic construction of an ontology and controlled vocabulary, as well as automatic mapping between tags and controlled vocabularies.

The results of our approach are encouraging for tag relation extraction. We plan to improve the classifier by collecting more relation types and human-coded examples for future experiments, and eventually utilize the relations extracted to enhance subject metadata descriptions.

References

Agichtein, Eugene, and Luis Gravano. (2000). Snowball: Extracting relations from large plain-text collections. In Kenneth M. Anderson, et al. (Ed.), *Proceedings of the 5th ACM Conference on Digital Libraries,* (pp. 85-94). New York: Association for Computing Machinery.

Brin, Sergey. (1998). Extracting patterns and relations from the World Wide Web. In Paolo Atzeni et al. (Ed.), *Selected Papers from the International Workshop on the World Wide Web and Databases,* (pp. 172-183). London: Springer.

Bunescu, Razvan C., and Raymond J. Mooney. (2007). Extracting relations from text from word sequences to dependency paths. In Anne Kao, et al. (Ed.), *Text Mining and Natural Language Processing,* (pp. 29-44). London: Springer.

Culotta, Aron, and Jeffrey Sorensen. (2004). Dependency tree kernels for relation extraction. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Retrieved April 13, 2008, from http://acl.ldc.upenn.edu/P/P04/P04-1054.pdf.

Culotta, Aron, Andrew McCallum, and Jonathan Betz. (2006). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics,* (pp. 296-303).

Guy, Marieke, and Emma Tonkin. (2006). Folksonomies: Tidying up tags? *D-Lib Magazine, 12*(1). Retrieved April 13, 2008, from http://www.dlib.org/dlib/january06/guy/01guy.html.

Heymann, Paul, and Hector Garcia-Molina. (2006). *Collaborative creation of communal hierarchical taxonomies in social tagging systems.* Technical Report 2006-10. Department of Computer Science, Stanford University. Retrieved April 13, 2008, from http://labs.rightnow.com/colloquium/papers/tag_hier_mining.pdf.

Iria, Jose, and Fabio Ciravegna. (2005). Relation extraction for mining the semantic web. *Dagstuhl Seminar on Machine Learning for the Semantic Web*. Retrieved April 13, 2008, from http://tyne.shef.ac.uk/t-rex/pdocs/dagstuhl.pdf.

Liu, Hugo and Pattie Maes. (2007). Introduction to the semantics of people & culture (Editorial preface). *International Journal on Semantic Web and Information Systems, Special Issue on Semantics of People and Culture, 3*(1). Retrieved March 28, 2008, from http://larifari.org/writing/IJSWIS2007-SPC-EditorialPreface.pdf.

Mathes, Adam. (2004). *Folksonomies-Cooperative classification and communication through shared metadata.* Unpublished manuscript. Retrieved April 13, 2008, from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

Mika, Peter. (2005). Ontologies are us: A unified model of social networks and semantics. In Yolanda Gil, et al. (Eds.), *Proceedings of the 4th International Semantic Web Conference (ISWC 2005),* (pp. 522–536). Berlin: Springer. Retrieved March 28, 2008, from http://ebi.seu.edu.cn/ISWC2005/papers/3729/37290522.pdf.

Michlmayr, Elke, Sabine Graf, Wolf Siberski, and Wolfgang Nejdl. (2005). A case study on emergent semantics in communities. In Yolanda Gil, et al. (Eds.), *Proceedings of the Workshop on Social Network Analysis, the 4th International Semantic Web Conference (ISWC 2005).* Berlin: Springer.

Nahm, Un Y., and Raymond J. Mooney. (2000). A mutually beneficial integration of data mining and information extraction. *Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence,* (pp. 627-632). Menlo Park, CA: AAAI Press.

Nguyen, Dat P., Yutaka Matsuo, and Mitsuru Ishizuka. (2007). Relation extraction from Wikipedia using subtree mining. *Proceedings of the National Conference on Artificial Intelligence Ontology Learning in conjunction with the 14th European Conference on Artificial Intelligence, Berlin, Germany*. Retrieved April 13, 2008, from http://acl.ldc.upenn.edu/N/N07/N07-2032.pdf.

Qin, Jian. (2008). Controlled semantics vs. social semantics: An epistemological analysis. *Proceedings of the 10th International ISKO Conference: Culture and Identity in Knowledge Organization, Montreal, 5.-8. August, 2008.* Retrieved March 28, 2008, from http://web.syr.edu/~jqin/pubs/isko2008_qin.pdf.

Rattenbury, Tye, Nathaniel Good, and Mor Naaman. (2007). Towards automatic extraction of event and place semantics from Flickr tags. In Charles L. Clarke, et al. (Ed.), *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* (pp. 103-110). New York: Association for Computing Machinery.

Roth, Dan, and Wen-tau Yih. (2002). Probabilistic reasoning for entity & relation recognition. *Proceedings of 19th International Conference on Computational Linguistics, 1-7*. New Brunswick: ACL.

Sanderson, Mark, and Bruce Croft. (1999). Deriving concept hierarchies from text. In M. Hearst, et al. (Ed.): *Proceedings of the 22nd ACM Conference of the Special Interest Group in Information Retrieval,* (pp. 206-213). New York: Association from Computing Machinery.

Schmitz, Patrick. (2006). Inducing Ontology from Flickr Tags. *Collaborative Web Tagging Workshop at WWW 2006*, *Edinburgh, UK*. Retrieved April 13, 2008, from http://www.topixa.com/www2006/22.pdf.

Shannon, Claude E. (1948). The mathematical theory of communication. *Bell System Technology Journal, 27,* 379-423.

Zelenko, Dmitry, Chinatsu Aone, and Anthony Richardella. (2003). Kernel methods for relation extraction. *Journal of Machine Learning Research, 3*, 1083-1106.