Automatic Generation of Metadata Elements to Textual Data[1]

Woojin Paik(UMass Boston)

1. Introduction

For a number of years both manual and automatic approaches to the construction of knowledge bases have been studied and implemented. Manual construction of knowledge bases has been too expensive to be practical and automatic approaches have not yet produced domain-independent and usable knowledge bases (Paik, 2000).

Lack of practically usable knowledge bases led to two key problems in preventing wide-scale deployment of knowledge-based systems; that is the knowledge base and inference engine. These problems are commonly referred to as brittleness and the knowledge acquisition bottleneck (Musen, 1989). A brittle system can respond appropriately only to a narrow range of questions. More precisely, such a system cannot answer questions that were not originally anticipated by the programmer. The other problem with knowledge-based systems is that crafting the statements that are entered into the knowledge base requires an enormous amount of training, time, and effort. Knowledge engineers tend

[1] Part of this research was supported by the National Science Foundation (NSF) grant number DUE-0085838.

to be highly skilled people but few of them can enter more than a small number of statements into a knowledge base in an average day. Brittleness and the knowledge-acquisition bottleneck are severe limitations. In recent years there has been increased interest in textual information extraction research using natural language processing techniques. The most common medium of storing knowledge is text; textual information extraction is an approach to acquire knowledge from text.

The study reported in this paper describes an adaptation of a Natural Language Processing (NLP) based information extraction system, which was originally developed to automatically populate knowledge bases, as a metadata extraction system for the digital libraries as well as a user preference elicitation tool.

2. Metadata Generation to Enable NSDL

Since the mid-1990s, there have been well-orchestrated movements to solve the general problem of networked information discovery and retrieval (NIDR) on the Internet. There is a growing consensus that in order for the emerging organizing systems for networked information, such as digital libraries, to interoperate, they must be based on some level of metadata standardization. Major national and international efforts are under way to create metadata element sets for cross-domain NIDR as seen in the Dublin Core Element Set (http://purl.oclc.org/dc) under development through the Dublin Core Metadata Initiative (DCMI).

Since 1995 the creation of education-specific metadata element sets to solve the general problem of Networked Information Discovery and Retrieval (NIDR) on the Internet has been the goal of both a number of government sponsored as well as private sector initiatives in the United States. The U.S. Department of Education's Gateway to Educational Materials (GEM) (http://www.TheGateway.org/) provides a good example of the former while the EDUCAUSE-sponsored Instructional Management System (IMS) project (http://www.imsproject.org/) is an excellent example of the latter. Use of these standards will enhance the understanding and sharing of data, information and processes to support, for

example, interoperability, electronic commerce, and component-based development of educational objects.

2. 1 GEM Meta-tags

The GEM meta-tag element set was the tag set that the system learned to assign automatically. The GEM element set uses the Dublin Core Element Set (DCES) as its base referent. In addition to the DCES fifteen elements and range of element qualifiers suited to general cross-domain networked information discovery and retrieval (NIDR), GEM adds eight education-specific elements and another range of element qualifiers. The table below defines the full element set (absent element qualifiers).

DUBLIN CORE ELEMENT	**LABEL**	**DESCRIPTION**
Title	Title	The name given the resource by the creator or publisher
Author or Creator	Creator	The person/organization primarily responsible for the intellectual content
Subjects and Keywords	Subject	The topic of the resource
Description	Description	A textual description of the content of the resource
Publisher	Publisher	The entity responsible for making the resource available in its present form
Other Contributor	Contributor	Secondary contributors to the intellectual content
Date	Date	The date the resource was made available in its present form
Resource Type	Type	The category of the resource
Format	Format	The data format of the resource
Resource Identifier	Identifier	A string or number that uniquely identifies the resource
Source	Source	A string or number used to uniquely identify the work from which this resource was derived
Language	Language	Language of the intellectual content of the resource
Relation	Relation	The relationship of this resource to other resources
Coverage	Coverage	The spatial and/or temporal characteristics of the resource
Rights Management	Rights	A link to copyright and/or use restriction statements
GEM ELEMENT	**LABEL**	**DESCRIPTION**
Audience	Audience	Information from a controlled vocabulary that most closely identifies the specific audience of the resource being described.

Duration	Duration	Recommended time or number of sessions needed to do the activity/lesson as stated in the entity being described
Cataloging	Cataloging	Information about the agency that created the GEM catalog record
EssentialResources	Resources	Free-text listing of materials essential to the successful use of the entity by the teacher as stated in the entity being described
EducationLevel	Level	Grade, grade span, educational level, or age of the entity's audience
Pedagogy	Pedogogy	Student instructional groupings, teaching methods, assessment methods, and learning prerequisites of a resource
Quality	Quality	Quality Indicators element is a means for assessing the quality of instructional materials
Standards	Standards	State, national, professional, or organizational standards mapped to the entity being described

An illustrative metadata record with information for select elements (and element qualifiers) looks like the following:

< DC.identifier> http://www.nytimes.com/learning/teachers/lessons/9808i2wednesday.html
<DC.type> Lesson plan
<DC.publisher> The New York Times Electronic Media Co.
<GEM.level> 6-12
<DC.subject> Social studies—World history
<DC.subject.keywords> Nuclear Non-Proliferation Treaty | Disarmament | Nuclear arms debate
<DC.Description> This lesson plan is designed to allow students to speak objectively about the nuclear disarmament issue and to interpret sections of the Nuclear Non-Proliferation Treaty. Students will become more informed by these discussions and readings on the nuclear arms debate and will thus be able to more adequately support any opinions they may have on the issue.
<GEM.standard> McREL
<GEM.standards.grade> 6-8
<GEM.standards.main> World History Standard 43 - Understands how post-World War II reconstruction occurred, new international power relations took shape, and colonial empires broke up.
<GEM.standards.subordinate> Understands post-war relations between the Soviet Union, Europe, and the United States

2.2 Automatically Assigning Educational Metadata

The applications described in this paper are an adaptation of <!metaMarker>, a commercially available metadata generation system (Paik & Brown, 2000). <!metaMarker> was initially designed to provide an "information context" in the form of a rich set of metadata tags for a variety of time and resource intensive tasks such as Customer Relation Management (CRM) and enterprise information filtering. <!metaMarker> automatically organizes customer service requests or incoming email streams according to their subject contents. It also automatically identifies such things as the emotional "tone" of the message and the intention or goal of the author of the message.

The underlying model of the processing algorithm behind the metadata extraction system is a recently emerged broad and shallow information extraction framework that was researched in the context of developing an information extraction system to automatically update knowledge bases (Paik, 2000). In comparison to the traditional deep and narrow information extraction systems such as the ones reported in the Message Understanding Conferences (MUC-3, 1991, MUC-4, 1992, MUC-5, 1993, & MUC-6, 1995) which require extensive manual development effort by subject matter experts, the broad and shallow information extraction systems are considered to more easily adaptable to new subject domains (Paik, 2000).

The core information extraction algorithm is based on sub-language analysis of text by taking advantage of the common practices of writers on a similar subject (Sager, et al, 1987). For example, there are regularities in the way that weather reports are composed. It is fairly straightforward to develop rules to extract key information about the weather reports by anticipating what type of information will be described in what manner. Similarly, previous work has shown that it is possible to develop a sub-language grammar to extract highly accurate information from news type stories. In conjunction with the use of case grammar type simple semantic relations such as 'agent', 'location', and 'cause', the use of sub-language grammar has been shown to enable extraction of practical, usable information from news type text.

This approach to extracting information has been tested and shown successful in the Defense Advanced Research Project Agency (DARPA)'s High Performance Knowledge Base (HPKB) program (Paik, 2000). The system developed for HPKB exhibited both high precision and high recall for information extraction tasks. This type of information algorithm has been incorporated into the commercial version of <!metaMarker>, an eXtensible Markup Language (XML)-based automatic metadata generation tool. <!metaMarker> extracts and classifies information objects from numerous types of business communications. The foundation of <!metaMarker> is built upon the richness and accuracy of Natural Language Processing (NLP) techniques and the adaptability and customization potential of Machine Learning (ML). It utilizes an expanded metadata framework developed for enterprise communications consisting of:

- Traditional descriptive, citation-like features: author, subject, time/date/place of creation
- Descriptive features unique to business communications: company/organization information, a specific order, named product features
- Additional situational or use aspects which provide critical contextual information: author's intention or goal, degree of certitude or conviction, mood or attitude

<!metaMarker> also facilitates addition of custom categories by derivation from previously extracted information. For example, extracted metadata elements such as 'subject', 'intention', and 'mood' might be used as the basis for defining another tag 'priority' that could be automatically assigned to a specific email based on the extracted values for the three original metadata elements. One possible instantiation is 'high' value assigned to 'priority' element if 'return of purchased product' was the value for 'subject' metadata element, 'complain' was the value for 'intention' element, and 'angry' was the value for 'mood' element.

In applying <!metaMarker> to email communication, derivation of relevant metadata elements was accomplished through both inductive means by analyzing a large number of emails, and deductive means by considering general theories of human communications and research results in the area of

computer mediated communication. There were some explicit metadata elements and their values were directly extractable from the body of email messages. For example, typical biographical information such as 'name of sender', 'title', 'affiliation', 'physical address', 'phone number', 'home page', or 'motto', were extracted by applying an email sublanguage grammar. The email sublanguage grammar was developed based on an analysis of output from various natural language processing components such as the 'proper name concept boundary identification and categorization module'.

There were also implicit metadata elements and their values identifiable through an email discourse model analysis. These elements were, 'subject/topic'. 'intention', and 'mood'. Subject/topic refers to the classification of the message contents into categories similar to those used in a general purpose thesaurus such as Roget's. Some examples of the values for this element are: law & politics, religion, science & technology, business & economics, and recreation & sports. The 'intention' metadata element comes from Searles' (1969) speech act theory, which focuses on what people 'do' with language, i.e. the various speech acts that are possible within a given language. <!metaMarker> utilizes discourse analysis of the email messages to classify authors' intentions into values such as 'promises', 'requests', or 'thanking'. The 'mood' element refers to the email authors' emotional state. The values for this element are: 'strongly negative', 'negative', 'neutral', and 'positive'.

To adapt <!metaMarker> to extract the educational materials specific metadata elements, the initial target elements from GEM and Dublin Core were categorized into three groups depending on how they will be extracted. Some elements such as 'author' or 'publisher' were directly extracted from the texts by applying educational material specific sublanguage grammar. Other elements such as 'quality' or 'relation', which are implicit in the texts, were derived through the discourse model analysis of the educational materials. There are some elements such as 'educational level', which can be both explicit and implicit in the texts. For these elements, direct extraction through sublanguage grammar analyzer was attempted first, then the discourse model analyzer was applied if the first attempt did not extract any values for the specific elements. The NLP part of the overall system consists of a number of core text processing components including a part-of-speech tagger, phrasal concept identifier, proper name

concept boundary identification & categorizer, and numeric concept boundary identification and categorizer.

3. Metadata Generation to Enable Personalization

In its most general form, personalization modifies an underlying system to better address the preferences of end users, be they corporate professionals or consumers (Smith, 2000). The Profile, which is the collection of data describing the criteria for customizing presentation or content, is the key to personalization. Linguistically speaking, personalization can be considered as a way to satisfy the Maxim of Relation (Grice, 1975). According to Grice, in a talk exchange the participants are expected to be conscious of the so-called Cooperative Principle, which states: "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (Grice, 1975). Conversing in accordance with the Cooperative Principle will yield maxims of Quantity (i.e. Don't say more or less than is required), Quality (i.e. Tell what you believe is true, be sincere), Relation (i.e. Be relevant), and Manner (i.e. Avoid ambiguity and obscurity) (Brown & Stephene, 1975).

On the other hand, personalization has different meanings to different people. Today, the three most common forms of personalization are: Enterprise-Controlled, End-user Controlled, and Data-Controlled (Votsch & Linden, 2000). The Enterprise-controlled form of personalization is making decisions based upon the preferences or predefined criteria set by the owner of the content. Criteria may be based on the factors of target platform, user role, level of service, or information extracted from an enterprise or a third-party repository. The systems of this type control access to content or functionality based on what the user is likely to purchase or has licensed. End-user controlled content delivery is based on criteria set by the customer. User controlled content applications in portals and in the enterprise context are examples of end-user controlled form of personalization (Smith, 2000 & Votsch & Linden, 2000). Data-controlled personalization is generated by affinity-data; for instance, the purchasing patterns and preferences of like consumer groups. Affinity-data are derived by applying

data-mining algorithms to market basket analysis. Affinities can be used to fine-tune customer interaction. For example, data-mining questionnaires can reveal the dislikes of different customer groups which can be further used to refine marketing campaigns. Furthermore, methods like collaborative filtering explore the choices of similar peer groups and recommend what other customers did at a certain point. Another form of data-controlled personalization is to leverage similarity of product descriptions in electronic product catalogs to cross-market similar products, given consumers' interest in a particular product (Votsch & Linden, 2000).

3. 1 Metadata Generation Example

The following is a sample email communication between a financial analyst and his/her client.

Question from a client:
I think the key to the future is the use of personalization software. Do you think BroadVision will rebound to its high in the next six months?

Response from a financial analyst:
BroadVision is more heavily concentrated in the B2B market, which, long term, we believe, is attractive. Though we like BroadVision, we think Ariba; I2 Technologies; and Commerce One will be the dominant players.

In addition to the typical metadata which are proper named concepts or numeric concepts, the user preference specific <!metaMarker> extracts the concepts that client liked, disliked, and also was interested in from this example. When the same type of information extraction is applied to the financial analyst's response, <!metaMarker> also extracts the concepts that the financial analyst's liked. In the following, a step-by-step analysis of the client question will be shown. This depiction shows the underlying NLP and ML processing of <!metaMarker>.

Step #1 (NLP) – sentence boundary identification
<s#1> I think the key to the future is the use of personalization software. </s#1> <s#2> Do you think BroadVision will rebound to its high in the next six months? </s#2>
<s> denotes the beginning of a sentence and </s> denotes the end of a sentence.

Step #2 (NLP) – part-of-speech tagging
<s#1> I|PRP think|VBP the|DT key|NN to|TO the|DT future|NN is|VBZ the|DT use|NN of|IN personalization|NN software|NN .|. </s#1> <s#2> Do|MD you|PRP think|VBP BroadVision|NP will|MD rebound|VB to|TO its|PRP$ high|JJ in|IN the|DT next|JJ six|CD months|NNS ?|. </s#2>
This step assigns part-of-speech information after each word in the sentence. '|' is used to delimit the word and the corresponding part-of-speech tag. The tag set is based on University of Pennsylvania's Penn Treebank Project (Santorini, 1990). For example, PRP means 'personal pronoun', VBP means 'present tense verb', and DT means 'determiner'.

Step #3 (NLP) – morphological analysis
<s#1> I|PRP think|VBP the|DT key|NN to|TO the|DT future|NN ***is|VBZ|be*** *the|DT use|NN of|IN personalization|NN software|NN .|. </s#1> <s#2> Do|MD you|PRP think|VBP BroadVision|NP will|MD rebound|VB to|TO its|PRP$ high|JJ in|IN the|DT next|JJ six|CD* ***months|NNS|month*** *?|. </s#2>*
This step determines the root form of each word and adds it to each word. In this example, there are two cases. 'is' is assigned with 'be' and 'months' is assigned with 'month'.

Step #4 (NLP) – multi-word concept identification
<s#1> I|PRP think|VBP the|DT key|NN to|TO the|DT future|NN is|VBZ|be the|DT use|NN of|IN ***<cn>*** *personalization|NN software|NN* ***</cn>*** *.|. </s#1> <s#2> Do|MD you|PRP think|VBP* ***<pn>*** *BroadVision|NP* ***</pn>*** *will|MD rebound|VB to|TO its|PRP$ high|JJ in|IN the|DT* ***<nc>*** *next|JJ six|CD months|NNS|month* ***</nc>*** *?|. </s#2>*
This step identifies the boundary of the concepts. For example, proper names are identified by <pn> tags. Numeric concepts are delimited by <nc> tags. All other multi-word concepts are bracketed by <cn> tags.

Step #5 (NLP) – concept categorization
<s#1> I|PRP think|VBP the|DT key|NN to|TO the|DT future|NN is|VBZ|be the|DT use|NN of|IN <cn> personalization|NN software|NN </cn> .|. </s#1> <s#2> Do|MD you|PRP think|VBP ***<pn cat=company>*** *BroadVision|NP </pn> will|MD rebound|VB to|TO its|PRP$ high|JJ in|IN the|DT* ***<nc cat=time>*** *next|JJ six|CD months|NNS|month </nc> ?|. </s#2>*

Each proper name and numeric concept is assigned with its semantic type information according to the predetermined schema. Currently, there are about 60 semantic types, which are automatically determined by <!metamarker>.

Step #6 (ML) – implicit metadata – mood, urgency, intention, and topic – generation
<s#1> I|PRP think|VBP the|DT key|NN to|TO the|DT future|NN is|VBZ|be the|DT use|NN of|IN <cn> personalization|NN software|NN </cn> .|.
<modalityInfo>
<mood> neutral </mood>
<urgency> neutral </urgency>
<intention> belief & judgment </intention>
</modalityInfo>
<topic> computer science & technology </topic>
</s#1>
<s#2>
Do|MD you|PRP think|VBP <pn cat=company> BroadVision|NP </pn> will|MD rebound|VB to|TO its|PRP$ high|JJ in|IN the|DT <nc cat=time> next|JJ six|CD months|NNS|month </nc> ?|.
<modalityInfo>
<mood> neutral </mood>
<urgency> neutral </urgency>
<intention> belief & judgment </intention>
</modalityInfo>
<topic> trade & commerce </topic>
</s#2>

This step assigns implicit metadata to each sentence by categorizing each sentence according to the predetermined schema of modality and topic/subject. The sentence-by-sentence categorization is carried out by the text classifiers such as Bayesian probabilistic classifier or k-Nearest Neighbor classifier by utilizing a training data set, which consists of a pre-coded set of example sentences. Each sentence is represented as a feature vector, which consists of NLP extracted explicit metadata from the steps #1 to #5. At the end of this stage, <!metaMarker>, which is not adapted to extract user preferences, is designed to generate a table to be incorporated as a part of a relational database.

Step #7 (ML) – user preference extraction
<s#1> I|PRP think|VBP the|DT key|NN to|TO the|DT future|NN is|VBZ|be the|DT use|NN of|IN <cn> personalization|NN software|NN </cn> .|.
<modalityInfo>

<mood> neutral </mood>
<urgency> neutral </urgency>
<intention> belief & judgment
<like> personalization software </like>
</intention>
</modalityInfo>
<topic> computer science & technology </topic>
</s#1>
<s#2>
Do|MD you|PRP think|VBP <pn cat=company> BroadVision|NP </pn> will|MD rebound|VB to|TO its|PRP$ high|JJ in|IN the|DT <nc cat=time> next|JJ six|CD months|NNS|month </nc> ?|.
<modalityInfo>
<mood> neutral </mood>
<urgency> neutral </urgency>
<intention> belief & judgment
<interested> BroadVision/company
</interested>
</intention>
</modalityInfo>
<topic> trade & commerce </topic>
</s#2>

Currently, the scope of the adaptation of <!metaMarker> to extract user preferences is limited to four types of metadata. They are 'like', 'dislike', 'interested', and 'not interested'. The user preference extraction is a combination of explicit and implicit metadata generation methods. First each sentence is categorized according to the positive and negative facets of 'like' and 'interested' user preferences. Then, certain explicit metadata extraction results such as proper names and multi-word concepts other than numeric concepts for each sentence is correlated with the user preference information. The above output of the step #7 shows that the client likes 'personalization software' and is interested in the company, BroadVision. This information will be entered into the user preference database so that the next interaction between the financial analyst and his/her client can be better focused on the clients' likes and interests. In addition, it is also expected that the financial analyst can push out certain relevant information to the client according to his/her preferences.

4. Evaluation

This paper is based on an ongoing project. Thus, the following table shows the partial experiment results for generating educational metadata. There were a total of 250 educational resources. Two-thirds of the resources were used for training and one-thirds for testing.

Two methods of measuring effectiveness that are widely used in the information extraction research community have been selected to evaluate the metadata extraction (Chincor, 1992). The methods are:

- **Precision**: the percentage of actual answers given that are correct.
- **Recall**: the percentage of possible answers that are correctly extracted.

Automatically extracted metadata was evaluated with the following criteria:

- If the automatically extracted metadata and the answer key, which is generated manually, are deemed to be equivalent, then the automatic extraction output is considered as "correct."
- If the automatically extracted information and the answer key do not match then it is considered as "incorrect."

Recall and precision are represented by the following equation (*possible* is defined as a sum of correctly extracted and missing metadata, and *actual* is defined as a sum of correctly extracted and incorrectly extracted metadata:

- *Recall = correct/possible*
- *Precision = correct/actual*

The following steps were followed to measure the effectiveness of automatically extracting metadata from educational resources:

- Test data was randomly selected and consisted of a pre-determined number of resources that were not used for training.
- A manual evaluation was conducted by presenting the automatically extracted metadata and the source text to three judges and asking them to categorize extracted metadata as correct or incorrect, and to identify missing information.
- Precision and recall were computed for the automatically extracted metadata by applying the majority principle (i.e. assume the correctness of a judgment if two or more judges make the same judgment.)
- A failure analysis was conducted of all incorrectly extracted missing information.

The following tables show the educational material specific metadata generation evaluation results.

Subject	Precision	Recall
Biological Sciences	96%	98%
Earth Sciences	93%	89%
Mathematics	100%	89%
Physical Sciences	91%	83%

Audience	Precision	Recall
Administrators	86%	99%
Bilingual Students	73%	99%
Elementary School Teachers	97%	93%
Female Students	100%	69%
Hearing-impaired Students	88%	99%
Hispanic-American Students	90%	100%
Middle School Teachers	100%	79%
Secondary School Teachers	91%	93%
Teachers	98%	92%
Parents	71%	95%

The metadata extraction experiment for the user preference extraction was conducted against 100 randomly selected customer inquiry email messages. The evaluation result for the user preference specific metadata using this previously unseen data is shown in the following table.

	Precision	Recall
Like	89%	85%
Dislike	91%	93%
Interested	88%	86%
Not Interested	82%	79%

It was expected that the 'Not Interested' category would result in the worst score since the development of the training data for this category was the most difficult one for the human coders. The humans had the most number of discrepancies for this category. On the contrary, 'Dislike' category scored best. This was also consistent with the human coders' experience with developing the training data set. They had the least discrepancies in finding email messages, which belong to the 'Dislike' category. The following table shows the Mood metadata element extraction evaluation result using the same 100 email messages.

	Precision	Recall
Positive	71%	81%
Neutral	90%	95%
Negative	93%	90%
Strongly Negative	86%	44%

The working definition of each category is developed inductively by analyzing the data. The 'Positive' category should be assigned when the customer is pleased with the transaction and openly expresses satisfaction and/or happiness. The 'Neutral' category means that the customer states fact or asks a question; does not express emotion either positively or negatively. The customer has found no fault with the service, web site, or product. The 'Negative' category should be assigned when the customer is dissatisfied with the transaction, and sometimes is openly negative, finding fault with the service, web site, or product and perhaps asking for clarification, explanation, or fix. The communication may include mild sarcasm. Finally, the 'Strongly Negative' means that the customer is

extremely dissatisfied with the transaction - disgusted, irate, and many times is going to cancel the order. This is communicated directly in the e-mail. Many times the e-mail shows caustic sarcasm.

We expected that if there is a small number of the training data for a certain category then the categorization effectiveness of that category is usually lower than the other categories with more training data. 'Positive' and 'Strongly Negative' categories had the lesser number of the training data in comparison to 'Negative' and 'Neutral' categories. The evaluation result confirms our hypothesis. It was also expected that there were high correlation between the occurrences of 'Positive' mood category with 'Like' and 'Interested' user preference categories. It turned out to be the case. In addition, 'Negative' and 'Strongly Negative' categories had high correlation with 'Dislike' category. However, the correlation between the negative mood categories and 'Not Interested' category had comparatively lower correlation. It seems that there are factors other than mood or emotions, which contribute to a customer not having interests in certain objects.

5. Conclusion

A combined NLP and ML approach to automate educational metadata and user preference extraction is introduced and its performance on a number of educational materials and email messages is described. The extended system, which is based on a general-purpose metadata generation system, accurately extracts various application specific metadata in addition to the traditional descriptive, citation-like features, descriptive features unique to business communications, and situational or use aspects. These metadata provide critical contextual information.

The same underlying metadata extraction framework that is implemented as <!metaMarker> is currently adapted for other application such as monitoring consumer perception of medical goods or services. The goal of this application is to monitor public perception of over-the-counter and prescription drugs. There are hundreds of chat rooms devoted to various medical conditions as well as discussion groups that discuss a particular medicine and its side effects. The proposed system will

automatically categorize harvested information according to the newly developed metadata elements such as *Condition, Side Effects, Severity of Side Effects, Off-label Use, Cures offered to Mitigate the Side Effects, Alternative Medicine, Source, Usage, and Attitude.*

The major potential contribution of the research reported in this paper is the demonstration of successfully using NLP and ML techniques as part of a large-scale work flow system to solve real-world problems. This success became possible due to the advancement of hybrid domain-independent and domain-dependent NLP techniques, which depart from the common practice of developing a specific one-off NLP application for each problem area.

6. References

Brown, P. & Stephen C.L. (1987). Politeness: Some universals in language usage. Cambridge: Cambridge University Press.

Grice, H.P. (1975). Logic & Conversation. Syntax & Semantics 3: 41-58.

MUC-3. (1991). Proceedings of the Third Message Understanding Conference (MUC-3), San Diego, CA, Morgan Kaufmann.

MUC-4. (1992). Proceedings of the Fourth Message Understanding Conference (MUC-4), McLean, VA, Morgan Kaufmann.

MUC-5. (1993). Proceedings of the Fifth Message Understanding Conference (MUC-5), Baltimore, MD, CA, Morgan Kaufmann.

MUC-6. (1995). Proceedings of the Sixth Message Understanding Conference (MUC-6), Columbia, MD, Morgan Kaufmann.

Musen, M.A. (1989). Widening the Knowledge-Acquisition Bottleneck: Automated Tools for Building and Extending Clinical Methods, in Hammond, W.E., Ed., AAAMSI Congress, San Francisco, CA.

Paik, W. (2000). CHronological information Extraction SyStem (CHESS), Ph.D. dissertation, Syracuse University, Syracuse, NY.

Paik, W. & Brown, E. (2000). Metadata: A Knowledge Management Enabler for Business Communication. Proceeding of Knowledge Management (KM) World 2000 Conference, Santa Clara, Ca.

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

Santorini, B. (1990). Part-of-speech Tagging Guidelines for the Penn Treebank Project. Technical report, Department of Computer & Information Science, U. of Penn.

Searl, J.R. (1969). Speech Acts: an Essay in the Philosophy of Language. Cambridge University Press. New York.

Smith, D. (2000). There Are Myriad Ways to Get Personal, Internet Week Online.

Votsch V. & Linden, A. (2000). Do you know what personalization means? Gartner Grp T-10-9346.