Recognizing Emotional Content of Emails as a byproduct of Natural Language Processing-based Metadata Extraction

이메일에 포함된 감성정보 관련 메타데이터 추출에 관한 연구

Woojin Paik(백우진)*

Abstract

This paper describes a metadata extraction technique based on natural language processing (NLP) which extracts personalized information from email communications between financial analysts and their clients. Personalized means connecting users with content in a personally meaningful way to create, grow, and retain online relationships. Personalization often results in the creation of user profiles that store individuals' preferences regarding goods or services offered by various e-commerce merchants. We developed an automatic metadata extraction system designed to process textual data such as emails, discussion group postings, or chat group transcriptions. The focus of this paper is the recognition of emotional contents such as mood and urgency, which are embedded in the business communications, as metadata.

초록

본 연구는 이메일에 나타난 감성정보 메타데이터 추출에 있어 자연언어처리에 기반한 방식을 적용하였다. 투자분석가와 고객 사이에 주고받은 이메일을 통하여 개인화 정보를 추출하였다. 개인화란 이용자에게 개인적으로 의미 있는 방식으로 컨텐츠를 제공함으로써 온라인 상에서 관계를 생성하고, 성장시키고, 지속시키는 것을 의미한다. 전자상거래나 온라인 상의 비즈니스 경우, 본 연구는 대량의 정보에서 개인에게 의미 있는 정보를 선별하여 개인화 서비스에 활용할 수 있도록, 이메일이나 토론게시판 게시물, 채팅기록 등의 텍스트를 자연언어처리 기법에 의하여 자동적으로 메타데이터를 추출할 수 있는 시스템을 구현하였다. 구현된 시스템은 온라인 비즈니스와 같이 커뮤니케이션이 중요하고, 상호 교환되는 메시지의 의도나 상대방의 감정을 파악하는 것이 중요한 경우에 그러한 감성정보 관련 메타데이터를 자동으로 추출하는 시도를 했다는 점에서 연구의 가치를 찾을 수 있다.

Keyword: Information Extraction, Metadata, Emotional Content

* Assistant Professor, Dept. of Computer Science, Konkuk University (wjpaik@kku.ac.kr)

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 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 user preference elicitation tool. The focus of this paper is the recognition of emotional contents embedded in the business communications by applying the information extraction technology to enable the data-controlled personalization in the context of e-business (Votsch & Linden 2000). Personalization modifies an underlying system to better address the preferences of end users, be they corporate professionals or consumers (Smith 2000). It often results in the creation of user profiles that store individuals' preferences regarding goods or services offered by various e-commerce merchants. The email communication between the financial analysts and their clients was selected as the source for extracting information to populate the client profiles. The personalization information extraction system was able to achieve a high level of accuracy.

2 Survey of Previous Works

There have been numerous studies on personalization especially in the e-commerce context. The meaning of personalization from various perspectives will be examined. Then, the roles of the emotions expressed in the written communication will be discussed. Finally, a number of automatic emotion recognition systems will be reviewed to gauge the state-of-art of the emotion recognition technology.

2.1 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. do not 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 & Stephen 1987).

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 controls 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 (Votsch & Linden 2000, Smith, 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).

2.2 Emotion

Fellous et al. (2004) described emotions as a difficult subject to study as the emotional experience is very personal and emotional reactions to situations vary greatly from person to person. Gill & Oberlander (2003) studied the relationship between language production and personality. They found that extraverts use wordier and more abstract language. They also found that extraverts' use of punctuations and

sentence structure were different from non-extraverts. Since high extraverts use more but less lexically specific words, there might be some differences between recognizing emotions from the texts written by high and low extraverts.

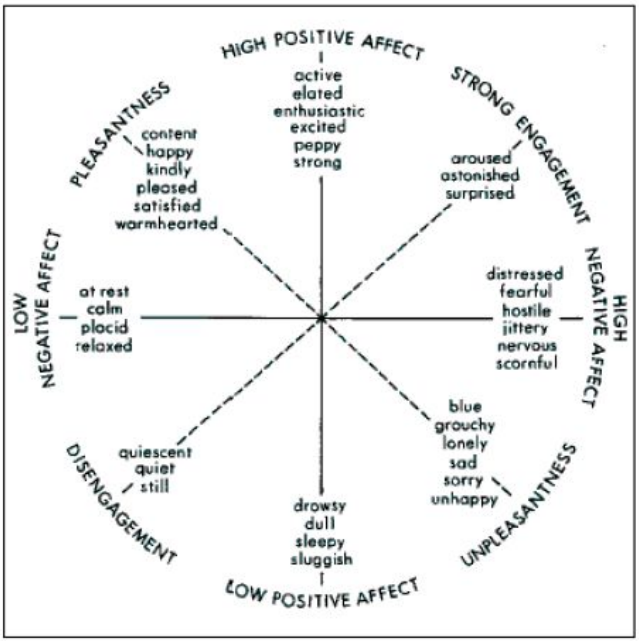

Figure 1: Two-dimensional Map of Affect. Reprinted from Watson and Tellegen (1985)

According to Ellis (2005), we do not know much about how people recognize emotions in text when there is not much context to rely on although that we know that language semantic especially vocabulary choice convey emotions. Within the typical text interactions such as email communications between people, each participant can expect about other person's normal language use, expected mood, and likely reactions to certain situations to a certain degree given their communication history. However, the business communications, namely email-based interactions between the customers and the customer representatives, differ from the previously described typical text interactions. There is probably little or no prior communication history. Nevertheless, this might make our task easier as the customers will most likely pay more attention in expressing what they want and how they feel in detail. This hypothesis will turn out to be correct if our experimental emotional content recognition rate is high.

There are many different emotional states including 1) moods such as cheerful, gloomy, irritable, listless, and depressed; 2) interpersonal stance such as distant, cold, warm, and supportive; 3) attitudes such as like, love, hate, value, and desire; and 4)

affect dispositions such as nervous, envious, reckless, morose, hostile (Ellis 2005). There is no consensus on which emotion set or emotion scale to use in categorizing emotions. Hernandez et al. (2004) used an emotion scale, which had five emotions such as happy, anger, sad, fear, and surprise. Watson & Tellegen (1985) studied the structure of affect, which resulted in the Circumplex theory of Affect. The theory utilizes two major bipolar dimensions of positive and negative affect. Positive affect signifies a combination of high energy and positive evaluation. Negative affect comprises feelings of upset and distress. Both positive and negative affect occur on bipolar continua, ranging from high to low. The Figure 1 depicts the two-dimensional map of affects. The eight octants shown in the Figure 1 were the emotion categories for the Natural Language Processing based text classification algorithm to assign to a given piece of written text (Rubin et al. 2004).

Our emotion scale was also influenced by the Circumplex theory of Affect. However, we only focused on the positive and negative bipolar continua instead of the actual emotion types. It resulted in a very simple scale with five categories such as strongly positive, positive, neutral, negative, and strongly negative. We considered these simple categories to be sufficiently distinctive in the context of prioritizing business emails from the customer service representatives' or the financial analysts' point of view.

2.3 Automatic Emotion Recognition

The most well known automatic emotion recognition system is an add-on module for a commercial email program. Kaufer (2000) discussed heuristics for identifying flaming text from web or emails. According to Wikipedia (Flaming), flaming is defined as "act of posting messages that are deliberately hostile and insulting." Eudora used some of these heuristics to develop a system to check outbound emails for flaming (MoodWatch). This program detects the sender's mood conveyed in the email messages by examining the words and sentences in the messages with respect to the entries in a flaming vocabulary/sentence dictionary. The program also examines the sentence structures. The program asks the sender if he/she really wants to send the email if the program determines that a particular email message exhibits flaming characteristics.

Lee et al. (2002) described methods to recognize emotions in spoken utterances, which were derived from telephone conversation, using acoustic and language information. To capture emotional information at the language level, they used emotional salience, which is defined as spot keywords. Linear discriminant classifiers were used for the acoustic information processing. To classify the emotions, k-nearest neighborhood classifiers were used. The combination of acoustic and language information to recognize negative emotions improved the accuracy by 45.7% from 32.9% accuracy achieved by the respective use of either acoustic or language information.

Conati & Zhou (2004) approached uncertainty in emotion recognition by relying on probabilistic reasoning, which is referred as Dynamic Bayesian Network. The

context of the study was the player affects where a socially intelligent agent interacts with an education game. There was no report on concrete emotion recognition accuracy.

To recognize emotions conveyed in sentences and/or discourses, a number of text analysis methods were developed. Initially, the words in the text segments are compared to the strong emotion bearing words in the pre-constructed keyword dictionaries. There were two major problems. One problem was the static nature of the dictionary where the emotion recognition wholly depended on the dictionary entries. The other problem was the dictionary based systems' inability to deal with the texts including negations (Ellis 2005).

Yu & Hatzivassiloglou (2003) conducted a research to separate opinions from facts. Firstly, they used Bayesian classifier to separate opinion news pieces such as editorials and fact-oriented regular news stories. Then, they used three increasingly sophisticated unsupervised statistical methods to detect sentences, which can be directly attributed as opinions. The methods were similarity approach, Naive Bayes Classifier, and multiple Naive Bayes Classifiers. Finally, they identified the polarity of each opinion sentences in terms of positive or negative with respect to the main perspective being expressed in the opinion sentence. The experimental result is based on 400 sentences. Binary classification of documents into either opinion or none-opinion achieved about 97% precision and recall. Finally, classifying opinion sentences by simple emotion scale such as positive, negative, or neutral achieved the accuracy close to 91%. This study is similar to our study in that they used a simple emotion scale and also with respect to the use of text classifier to do the emotion categorization.

3 System Description

One of the underlying text analysis models of the information extraction system described in this paper is a recently emerged broad & shallow information extraction framework. This domain-independent information extraction framework was used to develop an automated system to update knowledge bases (Paik 2000). In comparison to the traditional deep & narrow domain-dependent 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 the subject matter experts, the broad & shallow information extraction systems are considered to be more easily adaptable to new subject domains (Paik 2000).

3.1 Automatically Extracting User Preferences

Like many other systems, the domain-independent 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.

In this paper, we describe an eXtensible Markup Language (XML)-based automatic metadata generation system. It is a hybrid information extraction system, which utilizes both domain-independent and domain-dependent information extraction algorithms. The system does not extract case grammar type semantic relations like other information extraction systems. However, our system extracts and classifies information objects from numerous types of business communications. Our system is based on the Natural Language Processing (NLP) techniques and Machine Learning (ML). It utilizes an expanded metadata framework developed for enterprise communications consisting of: 1) traditional descriptive, citation-like features: author, subject, time/date/place of creation, 2) descriptive features unique to business communications: company/organization information, a specific order, named product features, and 3) additional situational or use aspects which provide critical contextual information: author's intention or goal, degree of urgency, mood or attitude.

It 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 our system 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 which were directly extractable from the body of email messages. For example, typical biographical information such as 'name of sender', 'title', 'affiliation', 'physical address', or 'phone number', 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 'concept categorization module'.

There were also implicit metadata elements and their values, identifiable through an email discourse model analysis. These elements were, 'subject/topic', 'intention', 'mood', and 'urgency'. Subject/topic refers to the classification of the message contents into categories such as are 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 Searl's speech act theory, which focuses on what people 'do' with language i.e. the various speech acts that are possible within a given language (Searl 1969). Our system utilizes discourse analysis of the email messages to classify authors' intentions into values such as 'claims', 'promises', 'requests', 'blessing', 'thanking', or 'permitting'. The 'mood' element refers to the email authors' emotional state. The values for this element are: 'strongly negative', 'negative', 'neutral', and 'positive'. Finally, 'urgency' is related to time, i.e. when something needs to be done (or was supposed to be done). The messages are classified and the following values are assigned to each message: 'very urgent', 'urgent', and 'neutral'.

In summary, our system is used as an implementation platform to automatically extract metadata for user preferences by incorporating user preferences specific extraction and tagging algorithms. To adapt our system to extract user preference specific metadata elements, the situational or use aspect related metadata are expanded to include new metadata elements such as 'like', 'dislike', 'interested', or 'not-interested.' Specifically these are the elements explaining the author's intention or goal. They are implicit in the text and thus derived through a text discourse model analysis of email type communicative text.

3.2 Applying Text Classifier to Extract Implicit Metadata

To assign implicit metadata to each sentence, each sentence is categorized according to the predetermined schema of modality and topic/subject. The first text classification task involves manually classifying a set of training documents in preparation for feeding the automatic system. Each training document is classified as "in" or "out" of the individual classes as outlined by the class definitions. The next step is to take these manually classified documents and process them through the trainable text classification system. During the process it builds a vector of terms, phrases, and entities extracted from the text. Multi-level Natural Language Processing outputs are the basis for these textual data feature representations. This collection of automatically generated features is then used to determine membership of new text within a particular class. The system determines the "certainty of membership" for each of the documents compared to each of the classes. If we consider a range of 1 to 0 where 1 means a document is definitely a member of a certain class, and 0 means a document is definitely a non-member of a certain class, we can say that values of 0 and 1 both have a "certainty of membership" value of 1. For either of these cases, we can confidently conclude that the document either 'does' or 'does not' belong within a given class. If we look at values close to .5 on the above scale, we have a "certainty of membership" value close to 0. This means for these cases, we cannot automatically determine whether or not a given document should be assigned to a given class. These documents are considered valuable in refining the classification system. By manually classifying these documents, and then feeding them back into the automatic system, we train it to recognize the subtle differences that distinguish how these documents should be classified.

3.3 NLP and ML Processing 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 the following, a step-by-step analysis of the client question will be shown. This depiction shows the underlying NLP and ML processing.

Step #1 (NLP) – sentence boundary identification (<s> denotes the beginning of a sentence and </s> denotes the end of a sentence.)

<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>

Step #2 (NLP) – part-of-speech tagging (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 [12]. For example, PRP means ' personal pronoun', VBP means 'present tense verb', and DT means 'determiner'.)

<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>

Step #3 (NLP) – morphological analysis (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'.)

<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>

Step #4 (NLP) – multi-word concept identification (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.)

<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>

Step #5 (NLP) – concept categorization (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 the NLP component of the system.)

<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>

Step #6 (ML) – implicit metadata – mood, urgency, intention, and topic generation (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 a Bayesian probabilistic text classifier through the use of 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.)

<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>

Step #7 (ML) – user preference extraction (Current system extracts four types of metadata about the user preferences. 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.)

<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>

4 Experiments and Results

Two methods of measuring effectiveness that are widely used in the information extraction research community have been selected to evaluate the metadata extraction including the user preference extraction performance (Chincor 1992). The methods are: 1) precision: the percentage of actual answers given that they are correct and 2) recall: the percentage of possible answers that are correctly extracted.

Automatically extracted metadata was evaluated with the following criteria: 1) 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" and 2) if the automatically extracted information and the answer key do not match then it is considered as "incorrect." Recall is the number of correct divided by number of possible. Precision is the number of correct divided by number of actual. The number of possible is defined as a sum of correctly extracted and miss-

ing metadata. The number of actual is defined as a sum of correctly extracted and incorrectly extracted metadata.

Explicit metadata extraction rules were developed inductively by analyzing randomly selected training data from a collection of actual emails which were sent by the customers of a commercial e-commerce merchant to the merchant. There were about 5,000 email messages in the training data set. The text classifier used to generate the implicit meta-data was trained by the same email messages after the appropriate implicit metadata including the user preferences was manually coded. The following steps were followed to measure the effectiveness of automatically extracting metadata from emails: 1) test data was randomly selected and consisted of a predetermined number of email messages that were not used for training, 2) 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, 3) 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), and 4) a failure analysis was conducted of all incorrectly extracted missing information.

The metadata extraction experiment was conducted against 100 randomly selected customer inquiry email messages. The evaluation result for the user preference specific meta-data using this previously unseen data is shown in the Table 1.

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

Table 1: User Preference Extraction Evaluation Results

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. Table 2 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

Table 2: Mood Extraction Evaluation Results

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. The customer is 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 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. Table 3 shows the Urgency metadata element extraction evaluation result using the same 100 email messages.

	Precision (%)	**Recall (%)**
Neutral	69	90
Urgent	82	85
Very Urgent	86	59
Urgent+Very Urgent	95	86

Table 3: Urgency Extraction Evaluation Results

The working definition of urgency is described in the following. The 'Neutral' category is assigned to the messages when they convey no sense of urgency. The 'Urgent' category means that the message conveys a need for action or response within a reasonable timeframe. However, no specific time needs to be mentioned. The 'Very Urgent' category means that the message conveys a need for an immediate action or response. Often times the action or response was desired or needed by the customer prior to writing the message. Finally, 'Urgent + Very Urgent' is used to categorize the messages at two dimensions namely that an urgency is conveyed in the message or not. We expected to see the better effectiveness when there is less number of categories for the system to learn. The evaluation results confirmed our

expectation. The decision to have more number of categories versus less number of categories for a certain metadata element is dependent on the application. The evaluation results shows one of the trade-offs of making such decision. There are five intention type metadata elements. They are: background, beliefs & judgments, niceties, promise, and request. The average precision of correctly assigning these intention metadata elements was 89.40% and the recall was 89.20%. The maximum precision value was 95% and the minimum precision value was 85%. The maximum recall value was 97% and the minimum recall value was 77%. These figures are based on the same experiment procedure described for measuring user preference type metadata element assignment effectiveness.

5 Conclusion

A combined NLP and ML approach to automate user preference extraction especially the emotional content is introduced and its extraction accuracy on a number of email messages is described. The extended system, which is based on a general-purpose metadata generation system, accurately extracts user preferences in addition to the traditional descriptive, citation-like features, descriptive features unique to business communication, and situational or use aspects which provide critical contextual information. Our system is designed to be a part of a larger Customer Relation Management (CRM) system that prioritizes & routes incoming customer inquiries and also populates user profiles. One possible extension of our system is the use of the developed technology in online reference system to improve the user satisfaction in electronically communicating with the reference librarians.

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 (e.g., CRM 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.

References

Chincor, N. 1992. MUC-4 Evaluation Metrics. Proceedings of the Fourth Message Understanding Conference, McLean, VA, USA.

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

Conati, C. & Zhou, X. 2004. A probabilistic framework for recognizing and affecting emotions. Proceedings of the AAAI 2004 Spring Symposium on Architectures for Modeling Emotions, Stanford University, CA.

Ellis, J.I. 2005. An exploration of human emotion perception from short texts. Artificial Intelligence and Psychology Project Report. School of Informatics, University of Edinburgh, UK.

Fellous, J.M. & Arbib, M.A. 2004. Emotions: From brain to robot. Trends in Cognitive Science, 8(12):pp.554-561.

Flaming. [Online] Available http://en.wikipedia.org/wiki/Flaming, April 30, 2006.

Gill, A. & Oberlander, J. 2003. Perception of e-mail personality at zeroacquaintance: Extraversion takes care of itself, but neuroticism is more of a worry. Proceedings of the 25th Annual Conference of the Cognitive Science Society, Boston, 2003. pp456–461.

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

Hernandez D.J., Deniz, O., Lorenzo, J. & Hernandez, M. (2004). BDIE: a BDI like architecture with emotional capabilities. Architectures for modeling emotion: Cross-Disciplinary foundations. Technical Report SS-04-02 of 2004 AAAI Spring Symposium.

Kaufer, D. 2000. Flaming: A white paper. Carnegie Mellon Univeristy.

Lee, C.M., Narayanan, S., & Pieraccuni, R. 2002. Combining acoustic and language information for emotion recognition. Proceedings of International Conference on Spoken Language Processing, Denver, CO. 873-876.

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

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

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

MUC-6 1995. Proceedings of the Sixth Message Understanding Conference. 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, USA.

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

MoodWatch. [Online] Available http://www.eudora.com/email/features/moodwatch.html, April 30, 2006.

Rubin, V.L., Stanton, J.M., & Liddy, E.D. 2004.Discerning emotions in texts. Proceeding of the 2004 AAAI Symposium on Exploring Attitude and Affect in Text, Stanford, CA, USA.

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

Searl, J.R. 1969. Speech Acts: an Essay in the Philosophy of Language. Cambridge University Press. NY, USA.

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 Group T-10-9346.

Watson, D. & Tellegen, A. 1985. Toward a consensual structure of mood. Psychological Bulletin, 98, 219-235.

Yu, H. & Hatzivassiloglou, V. 2003. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. Proceedings of the conference on Empirical Methods in Natural Language Processing.