# From whom, about what, and reason why

## Capturing public's perception using Natural Language Processing (NLP)

**Woojin Paik**

Dept. of Computer Science
University of Massachusetts Boston
100 Morrissey Blvd
Boston, MA 02125, USA
wjpaik@cs.umb.edu

**Jee Yeon Lee**

Dept. of Library & Information Science
Yonsei University
134 Sinchon-dong, Seodaemun-gu
Seoul 120-749, South Korea
jlee01@lis.yonsei.ac.kr

## Introduction

We describe a computational linguistics analysis method to capture public's perception about goods, services, and events, which are expressed in the forums such as the Internet chat rooms and the discussion groups.

The web has created an unprecedented opportunity to mine and organize the general population's experiences and opinions. There are thousands of chat rooms and discussion groups devoted to almost an infinite variety of topics including the world events such as the September 11th tragedy and the bio-terrorist attack and its antidote such as Cipro.

In dealing with various crisis situations or other mundane events such as an introduction of a new medicine, it is important to understand the changing nature of the public's perception to ensure that proper and relevant information is communicated amongst the public. This monitoring functionality can ensure the public's continued support for a just cause. It is also important to assess the impact of potential misinformation, which is injected intentionally or unintentionally by the various interest bearing parties to delineate their intentions and their potential future activities.

The quantitative survey method has been widely used to gauge the public perception. However, it is often costly and time consuming to develop the surveys, to gather the data, and to analyze the results. It is also difficult to get the timely assessment of the public's perception. It is due to the significant lapse between the time when the survey takes place and the time when the analysis results become available. More importantly, it is often not possible to ask certain types of questions to the public directly by using the survey methods. It is partly because that people will not tolerate long and time-consuming surveys. Another reason is that certain questions will be considered to be an invasion of privacy. In addition, certain questions will be just inappropriate to ask by using the survey method.

It will be easier to discover the important trends and develop effective and fast reactive measures to prevent or promote certain perceptions if it is possible to continuously monitor public's perception without human intervention. A prototype automatic and non-invasive public perception monitoring system incorporates various Natural Language Processing (NLP) techniques, including domain-independent information extraction, performative information capture, and text discourse analysis. The NLP system is designed to automatically analyze the textual data by extracting salient information and discovering underlying intentions of the message composers.

The prototype system can be considered as an autonomous agent, which scours the web to gather data from the public chat rooms, discussion groups, and message boards without interfering the regular activities. The agent monitors the web, develops the new public perception summaries, and augments the existing ones. The integration of the various Natural Language Processing (NLP) modules ensures the correct interpretation of the public views described in the textual communicative messages.

The resulting summaries are intended to enable the interested parties such as policy-makers, law enforcement, and even the general public to access and further interpret the vast amount of the complex web data efficiently. Ultimately, it is expected that the policy makers will be able to clear-up misperception about events by detecting unexpected public perception about those events. The law enforcement might be able to discover

new threats or clues to solving the existing mysteries by monitoring the public perception summaries.

In a manner similar to how the current survey data is used by many news agencies, it might be also possible for the public to view certain portions of the resulting summaries so that they can study independently about a particular event and make an informed decision about how they should react regarding the rumors or other people's opinions.

## Performative Information

To capture public perception, which is implied in the chat room or discussion group postings, it is important to extract what people 'do' with language in addition to the linguistic information embedded in the textual data. The speech act theory (Searle, 1969) is the basis for deciphering this performative information. Searle identified four types of speech acts. They are 1) utterances, 2) propositional utterances, 3) illocutionary utterances, and 4) perlocutionary utterances.

An utterance is a spoken word or string of spoken words. At the simplest level, to utter is simply to say a word with no particular forethought or intention to communicate a meaning. In this case, there is no intention to communicate meaning by an utterance. A propositional utterance is a more meaningful type of utterance, which makes reference to or describes a real or imaginary object. In the act of making a propositional utterance, the speaker gains the opportunity to interact. The propositional utterances need not be sentences, and they do not have to intend anything. An illocutionary utterance is spoken with the intention of making contact with a listener. Illocutionary utterances are usually sentences that contain propositional utterances, that is, they refer to things in the world. However, it is their intentional nature that is of the most importance. Once it becomes clear that the speaker's intention is important to the meaning of an utterance, it can be seen that the same set of words might have different meanings depending on the speaker's intention. A given utterance might become a statement, a command, a question, a request, a promise, and so on depending upon the context and the speaker's intention. Illocutionary speech acts may by intended to provide information, solicit answers to the questions, give praise, and so on, but they do not necessarily require that the listener change his or her behavior. Perlocutionary utterances, on the other hand attempt to effect a change. As with the others, perlocutionary speech acts are utterances; they include propositions, and they intend interaction with the receiver.

To get the full picture of the public perceptions about a certain event or a product/service, it is necessary to understand what each chat group participant or the discussion group posting writer's intention behind the words and phrases used to compose the messages are. The proposed public perception monitoring approach is designed to capture publics' intentions, which can be considered as the illocutionary and perlocutionary speech acts. A new processing module was developed to extract the publics' intentions. The new module was incorporated as a part of the overall NLP system. The new module is comprised of two sub-modules. One sub-module will convert the domain-independent information extraction output such as the POS tags, numeric concepts & their semantic categories, proper names & and their associated categories, and the predicate-argument pairs into a feature vector. Each sentence in the chat room or the discussion group posting has its own feature vector. The other sub-module is a text classifier, which takes the feature vector generated from the first module as the input, and then the classifier categorizes each sentence according to the different speech acts. For example, each sentence was categorized as one of the following four major types of communicative illocutionary acts. Then, each sentence were further categorized as one of many minor classes under each major type (Bach, 1994.)

1) Constatives
2) Directives
3) Commissives
4) Acknowledgments

## Text discourse structure based modeling

Sub-language grammar (Sager et al, 1987) reveals the grammatical regularities, which exist within a particular text type. The information extraction takes advantage of these regularities to make is operation domain-independent. In addition, there are organizational regularities, which exist within a text type. It is called discourse structure. A typical information organization scheme employed in a chat room discussion about various products including

medicine such as Cipro, which is used to treat Anthrax, reveals a cyclical model of how public expresses their views (figure 1). Both linguistic and performative information extraction output were organized according to this model. The text discourse structure was determined by the use of the base information extraction results as the input feature vectors for a text classifier.

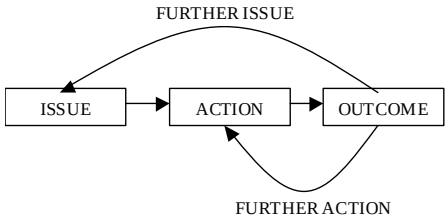


Figure 1. Cyclical Perception Model

Another major functionality of the NLP system is the subject domain specific semantic relation extraction algorithm. A set of semantic relations for a particular domain was used to anticipate the types of information to extract. For example, the semantic relations listed in the following were identified to adequately explain the chat room or the discussion group postings about the medical products. It is possible to develop the specific relations and their associated information extraction sub-language rules automatically given that there is enough number of examples for each semantic relation to model. The training examples in conjunction with the appropriate inductive machine-learning algorithm have shown to be able to generate the semantic relation extraction rules (Riloff, et al, 1998.)

1) Condition: The illness or health problem of the person
2) Side Effects: The side effects mentioned in the message
3) Severity of Side Effects: Major or Minor
4) Off-Label Use: The medication is used to treat another illness that isn't mentioned by the company.
5) Another Cause Mentioned for Side Effects: Other conditions or medicines that might have caused the side effects.
6) Cures Offered to Mitigate the Side Effects
7) Alternative Medicine: A medicine mentioned in the text that doesn't have the side effects.
8) Request for Information
9) Source: Source for the information provided in the text, i.e. doctor, speaker, nurse
10) Usage: How much was taken, for how long.
11) Attitude: The attitude of the person to the medicine mentioned.

There are a number of different attributes, which are associated with 'outcome', which is a component in the cyclical chat room discourse model. The attributes are: 'major', 'minor', 'positive', 'negative', 'intended', and 'not intended'. Additionally, 'attitude', 'source', and 'usage' can be attributes of the 'issue', 'action' and the 'outcome'. The following is a hypothetical discussion group posting.

*I went on Cipro last week to make sure that I do not get Anthrax. I heard that it was the best tool to prevent the infection. However, it made me extremely agitated, which frightened me, so I went off of it. Does anyone know of another drug that is just as good as Cipro, but won't turn me into a witch?*

The NLP system processed the example according to the following steps.

**Step #1 – sentence boundary identification**

*<s#1> I went on Cipro last week to make sure that I do not get Anthrax. </s#1> <s#2> I heard that it was the best tool to prevent the infection. </s#2> <s#3> However, it made me extremely agitated, which frightened me, so I went off of it. </s#3> <s#4> Does anyone know of another drug that is just as good as Cipro, but won't turn me into a witch? </s#4>*

<s> denotes the beginning of a sentence and </s> denotes the end of a sentence.

**Step #2 – part-of-speech tagging**

*<s#1> I|PRP went|VBD on|IN Cipro|NP last|JJ week|NN to|TO make|VB sure|JJ that|IN I|PRP do|VBP not|RB get|VB Anthrax|NP .|. </s#1> <s#2> I|PRP heard|VBD that|IN it|PRP was|VBD the|DT best|JJS tool|NN to|TO prevent|VB the|DT infection|NN .|. </s#2> <s#3> However|RB ,|, it|PRP made|VBD me|PRP extremely|RB agitated|VBD ,|, which|WP frightened|VBD me|PRP ,|, so|IN I|PRP went|VBD off|IN of|IN it|PRP .|. </s#3> <s#4> Does|VBZ anyone|NN know|VB of|IN another|DT drug|NN that|WDT is|VBZ just|RB as|RB good|JJ as|IN Cipro|NP ,|, but|CC will|MD not|MD turn|VB me|PRP into|IN a|DT witch|NN ?|? </s#4>*

In this step, each word is assigned with a part-of-speech tag. '|' is used to delimit the word and the corresponding part-of-speech. 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 – morphological analysis**

*<s#1> I|PRP **went|VBD|go** on|IN Cipro|NP last|JJ week|NN to|TO make|VB sure|JJ that|IN I|PRP do|VBP not|RB get|VB Anthrax|NP .|. </s#1> <s#2> I|PRP **heard|VBD|hear** that|IN it|PRP **was|VBD|be** the|DT best|JJS tool|NN to|TO prevent|VB the|DT infection|NN .|. </s#2> <s#3> However|RB ,|, it|PRP **made|VBD|make** me|PRP **extremely|RB|extreme agitated|VBD|agitate** ,|, which|WP **frightened|VBD|frighten** me|PRP ,|, so|IN I|PRP **went|VBD|go** off|IN of|IN it|PRP .|. </s#3> <s#4> Does|VBZ anyone|NN know|VB of|IN another|DT drug|NN that|WDT **is|VBZ|be** just|RB as|RB good|JJ as|IN Cipro|NP ,|, but|CC will|MD not|MD turn|VB me|PRP into|IN a|DT witch|NN ?|? </s#4>*

This step determines the base form of each word and adds it to each word. For example, a past tense verb such as 'made' is normalized as 'make'.

**Step #4 – multi-word concept identification**

*<s#1> I|PRP went|VBD|go on|IN **<pn>** Cipro|NP **</pn>** **<nc>** last|JJ week|NN **</nc>** to|TO make|VB sure|JJ that|IN I|PRP do|VBP not|RB get|VB **<pn>** Anthrax|NP **</pn>** .|. </s#1> <s#2> I|PRP heard|VBD|hear that|IN it|PRP was|VBD|be the|DT **<cn>** best|JJS tool|NN **</cn>** to|TO prevent|VB the|DT infection|NN .|. </s#2> <s#3> However|RB ,|, it|PRP made|VBD|make me|PRP extremely|RB|extreme agitated|VBD|agitate ,|, which|WP frightened|VBD|frighten me|PRP ,|, so|IN I|PRP went|VBD|go off|IN of|IN it|PRP .|. </s#3> <s#4> Does|VBZ anyone|NN know|VB of|IN another|DT drug|NN that|WDT is|VBZ|be just|RB as|RB good|JJ as|IN **<pn>** Cipro|NP **</pn>** ,|, but|CC will|MD not|MD turn|VB me|PRP into|IN a|DT witch|NN ?|? </s#4>*

This step identifies the boundary of the concepts. For example, the <pn> tags identify the proper names. Numeric concepts are delimited by <nc> tags. All other multi-word noun compound concepts are bracketed by <cn> tags.

**Step #5 – concept categorization**

*<s#1> I|PRP went|VBD|go on|IN <**pn cat=drug**> Cipro|NP </pn> <**nc cat=time**> last|JJ week|NN </nc> to|TO make|VB sure|JJ that|IN I|PRP do|VBP not|RB get|VB <**pn cat=disease**> Anthrax|NP </pn> .|. </s#1> <s#2> I|PRP heard|VBD|hear that|IN it|PRP was|VBD|be the|DT <cn> best|JJS tool|NN </cn> to|TO prevent|VB the|DT infection|NN .|. </s#2> <s#3> However|RB ,|, it|PRP made|VBD|make me|PRP extremely|RB|extreme agitated|VBD|agitate ,|, which|WP frightened|VBD|frighten me|PRP ,|, so|IN I|PRP went|VBD|go off|IN of|IN it|PRP .|. </s#3> <s#4> Does|VBZ anyone|NN know|VB of|IN another|DT drug|NN that|WDT is|VBZ|be just|RB as|RB good|JJ as|IN <**pn cat=drug**> Cipro|NP </pn> ,|, but|CC will|MD not|MD turn|VB me|PRP into|IN a|DT witch|NN ?|? </s#4>*

Each proper name and numeric concept is assigned with its semantic type information according to the predetermined schema. The categorization is based on about 60 classes. For example, Cipro is categorized as a drug name.

**Step #6 – predicate-argument identification**

*<s#1> I|PRP **<pa>** went|VBD|go on|IN <pn cat=drug> Cipro|NP </pn> **<pa>** <nc cat=time> last|JJ week|NN </nc> to|TO make|VB sure|JJ that|IN I|PRP **<pa>** do|VBP not|RB get|VB <pn cat=disease> Anthrax|NP </pn> **</pa>** .|. </s#1> <s#2> I|PRP heard|VBD|hear that|IN it|PRP was|VBD|be the|DT <cn> best|JJS tool|NN </cn> to|TO **<pa>** prevent|VB the|DT infection|NN **</pa>** .|. </s#2> <s#3> However|RB ,|, it|PRP **<pa>** made|VBD|make me|PRP extremely|RB|extreme agitated|VBD|agitate **</pa>** ,|, which|WP <**pa**> frightened|VBD|frighten me|PRP </**pa**> ,|, so|IN I|PRP <pa> went|VBD|go off|IN of|IN it|PRP </pa> .|. </s#3> <s#4> Does|VBZ anyone|NN know|VB of|IN another|DT drug|NN that|WDT is|VBZ|be just|RB as|RB good|JJ as|IN <pn cat=drug> Cipro|NP </pn> ,|, but|CC **<pa>** will|MD not|MD turn|VB me|PRP into|IN a|DT witch|NN **</pa>** ?|? </s#4>*

This step identifies the predicate-arguments. In this application, the predicate-arguments are

defined as a verb followed by indirect/direct objects. '<pa>' and '</pa>' tags signal the beginning and end of a predicate argument. For example, the following three adjacent words, 'went on Cipro' are identified as a predicate-argument.

**Step #7 – issue, action, outcome, further action identification**

*<s#1> I|PRP* ***<action>*** *<pa> went|VBD|go on|IN <pn cat=drug> Cipro|NP </pn> <pa>* ***</action>*** *<nc cat=time> last|JJ week|NN </nc> to|TO make|VB sure|JJ that|IN I|PRP* ***<issue>*** *<pa> do|VBP not|RB get|VB <pn cat=disease> Anthrax|NP </pn>* ***</issue>*** *</pa> .|. </s#1> <s#2> I|PRP heard|VBD|hear that|IN it|PRP was|VBD|be the|DT <cn> best|JJS tool|NN </cn> to|TO* ***<issue>*** *<pa> prevent|VB the|DT infection|NN </pa>* ***</issue>*** *.|. </s#2> <s#3> However|RB ,|, it|PRP <**outcome**> <pa> made|VBD|make me|PRP extremely|RB|extreme agitated|VBD|agitate </pa> </**outcome**> ,|, which|WP <**outcome**> <pa> frightened|VBD|frighten me|PRP </pa> </**outcome**> ,|, so|IN I|PRP <**outcome**> <pa> went|VBD|go off|IN of|IN it|PRP </pa> </**outcome**> .|. </s#3> <s#4> Does|VBZ anyone|NN know|VB of|IN <**further_action**> another|DT drug|NN </**further_action**> that|WDT is|VBZ|be just|RB as|RB good|JJ as|IN <pn cat=drug> Cipro|NP </pn> ,|, but|CC <pa> will|MD not|MD turn|VB me|PRP into|IN a|DT witch|NN </pa> ?|? </s#4>*

This step classifies each predicate-argument, multi-word concept, or single-word concept according to one of the semantic relation types. The types used in this example are: 'issue', 'action', 'outcome', and 'further action'.

**Step #8 – implicit attribute (speech acts) identification**

*<s#1* ***mood=neutral intended=yes severity=minor*** *> I|PRP <action> <pa> went|VBD|go on|IN <pn cat=drug> Cipro|NP </pn> <pa> </action> <nc cat=time> last|JJ week|NN </nc> to|TO make|VB sure|JJ that|IN I|PRP <issue> <pa> do|VBP not|RB get|VB <pn cat=disease> Anthrax|NP </pn> </issue> </pa> .|. </s#1> <s#2* ***mood=neutral intended=no severity=minor*** *> I|PRP heard|VBD|hear that|IN it|PRP was|VBD|be the|DT <cn> best|JJS tool|NN </cn> to|TO <issue> <pa> prevent|VB the|DT infection|NN </pa> </issue> .|. </s#2> <s#3* ***mood=negative intended=no severity=major*** *> However|RB ,|, it|PRP <outcome> <pa> made|VBD|make me|PRP extremely|RB|extreme agitated|VBD|agitate </pa> </outcome> ,|, which|WP <outcome> <pa> frightened|VBD|frighten me|PRP </pa> </outcome> ,|, so|IN I|PRP <outcome> <pa> went|VBD|go off|IN of|IN it|PRP </pa> </outcome> .|. </s#3> <s#4* ***mood=neutral intended=yes severity=minor*** *> Does|VBZ anyone|NN know|VB of|IN <further_action> another|DT drug|NN </further_action> that|WDT is|VBZ|be just|RB as|RB good|JJ as|IN <pn cat=drug> Cipro|NP </pn> ,|, but|CC <pa> will|MD not|MD turn|VB me|PRP into|IN a|DT witch|NN </pa> ?|? </s#4>*

This step identifies the attributes of each semantic relation. In this application, there are three attributes – 'mood', 'intended', and 'severity'. These attributes are assigned to each sentence then assigned to each semantic relation.

The performative information extraction output is shown in the following table.

| | **Issue** | **Action** | **Outcome** | **Further Action** |
|---|---|---|---|---|
| | do not get Anthrax (disease); prevent infection | go on Cipro (drug) | make one agitated; frighten one; go off of it | another drug |
| **Mood** | neutral; | neutral | negative | neutral |
| **Intended** | Yes | yes | no | yes |
| **Severity** | major | major | major | minor |

## Conclusion

This is an ongoing research. We do not yet have concrete evaluation result to report. However, we are reasonably confident in achieving high accuracy and the coverage of extracting performative information based on our previous studies of extracting semantic relations in different domains (Paik et al, 2001, Paik, 2001, Paik, 2000.)

Intuitively, it is better to know the intention of the speaker in deducing what he/she meant. Thus, we argue that the extraction of the factual information alone from the communicative text is not enough to enable the effective intent inference. In this paper, we described a multi-stage NLP system to capture this subtle but significant piece of performative information

from Internet chat group postings, which is becoming one of the most popular ways to communicate.

## References


Bach, K. 1994. *Conversational implicature*, Mind & Language 9: 124-62.

Paik, W., Harwell, S., Yilmazel, S., Brown, E., Poulin, M., Dubon, S., & Amice, C. 2001. *Applying Natural Language Processing Based Metadata Extraction to Automatically Acquire User Preferences*. Proceedings of the First International Conference on Knowledge Capture. Victoria, British Columbia, Canada. Oct 21, 2001.

Paik, Woojin 2001. *Automatic Generation of Educational Metadata Elements to Enable Digital Libraries.* Proceedings of the International Conference on Computers in Education (ICCE) 2001. Seoul, Korea. Nov. 12, 2001.

Paik, Woojin 2000. *Chronological Information Extraction System*, Doctoral Dissertation, Syracuse University, Syracuse, NY.

Riloff, E. & Schmelzenbach, M. 1998. *An Empirical Approach to Conceptual Case Frame* Acquisition Proceedings of the Sixth Workshop on Very Large Corpora.

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

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

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