Monitoring Public Perception of Medical Products by Automatically Assigning Metadata to Online Discussion Group Postings

Woojin Paik, Sarah Harwell, Sibel Yilmazel, Eric Brown, Maryjane Poulin, Stephane Dubon, Christophe Amice
solutions-united, inc.
Syracuse, NY, USA

Abstract *This paper describes an ongoing project whose goal is to create a digital library of public perceptions of over-the-counter and prescription drugs. The Web has created an unprecedented opportunity to mine and organize public perceptions and experiences with medications. There are hundreds of chat rooms devoted to medical conditions and discussion groups for medicines and their side effects. Gathering information and discerning patterns by trawling through the web manually is an arduous and time-consuming project, both inefficient and incomplete. The system described in this paper automatically collects data from chat rooms and discussion groups; meta-tags the information and organizes the resulting meta-tags into a meaningful format. The technology is expected to enable manufacturers, medical researchers, regulators, and the public to access and interpret the data.*

Keywords: Natural Language Processing, Metadata, Perception Monitoring

1 Introduction

We are currently developing a public perception monitoring system using Natural Language Processing (NLP) and Machine Learning (ML). This development project aims to create a digital library of public perceptions of over-the counter and prescription drugs. The Web has created an unprecedented opportunity to mine and organize the general population's experiences with medications. 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. Gathering information and discerning patterns by trawling through the web manually is an arduous and time-consuming project, both inefficient and incomplete. The system in this ongoing project automatically collects data from medical chat rooms and discussion groups; automatically categorizes the information by meta-tagging it, and organizes the resulting meta-tags into a meaningful format. The resulting system will enable manufacturers, medical researchers, regulators, and the general public to access and interpret the data. Manufacturers will be able to clear up misperceptions about their products, identify potential problems not seen during initial testing, and monitor usage. Researchers and regulators will be able to easily and quickly uncover trends in side effects and the public will be able to research a particular drug to attain an informed decision about its use. It is expected that the resulting system will be available through a digital library and be interactive; allowing interested parties to post corrections, additions, and bulletins.

2 Project Description

The public perception monitoring system is based on <!metaMarker>, an automatic metadata extraction system based on NLP and ML techniques. <!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 [7]. In comparison to the traditional deep and narrow information extraction systems such as the ones reported in the Message Understanding Conferences [3,4,5,6] 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 [7].

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 [7]. 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 [7]. 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 with 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's [10] 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'.

In the research reported in this paper, <!metaMarker> was used as an implementation platform to automatically extract metadata to gather public perceptions about medical products by incorporating perception-description specific extraction and tagging algorithms. Specifically these are the elements explaining the author's description of his/her perception. They are implicit in the text and thus derived through a text discourse model analysis of discussion group postings, a type of communicative text. To adapt <!metaMarker> to extract public perception specific metadata elements, the topic/subject related metadata were expanded to include new metadata elements such as the following:

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

These categories are used to generate output that takes the cyclical model shown in the Figure 1. This model shows the main characteristic of the discourse structure of chat room talk.

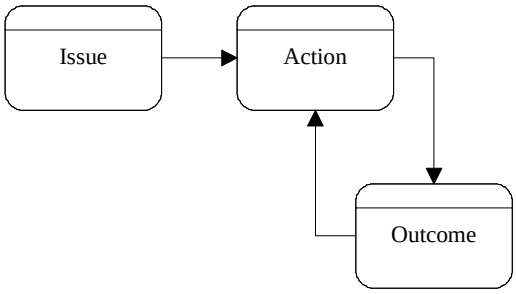

Figure 1: Cyclical Discourse Model

There are a number of different attributes associated with *outcome*. They 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 example shows how the model can be used to create meaningful content. Step-by-step analysis of the discussion group posting is shown. This depiction presents the underlying NLP and ML processing of <!metaMarker>.

Example Posting (input)
I went on Zyban in July to stop smoking and it was an incredible tool to keep away the cravings. However, it made me extremely agitated.

Step #1 (NLP) – sentence boundary identification
<s#1> I went on Zyban in July to stop smoking and it was an incredible tool to keep away the cravings. </s#1> <s#2> However, it made me extremely agitated. </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 went|VBD on|IN Zyban|NP in|IN July|NP to|TO stop|VB smoking|VBG and|CC it|PRP was|VBD an|DT incredible|JJ tool|NN to|TO keep|VB away|RB the|DT cravings|NNS .|. </s#1> <s#2> However|RB ,|, it|PRP made|VBD me|PRP extremely|RB agitated|VBD .|. </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 [9]. For example, PRP means 'personal pronoun', VBP means 'present tense verb', and DT means 'determiner'.

Step #3 (NLP) – morphological analysis
<s#1> I|PRP ***went|VBD|go*** *on|IN Zyban|NP in|IN July|NP to|TO stop|VB* ***smoking|VBG|smoke*** *and|CC it|PRP* ***was|VBD|be*** *an|DT incredible|JJ tool|NN to|TO keep|VB away|RB the|DT* ***cravings|NNS|craving*** *.|. </s#1> <s#2> However|RB ,|, it|PRP* ***made|VBD|make*** *me|PRP* ***extremely|RB|extreme agitated|VBD|agitate*** *.|. </s#2>*

This step determines the root form of each word and adds it to each word. For example, in the sentence above 'went' is assigned with 'go' and 'cravings' is assigned with 'craving'.

Step #4 (NLP) – proper name & multi-word concept identification
<s#1> I|PRP went|VBD|go on|IN ***<pn>*** *Zyban|NP* ***</pn>*** *in|IN* ***<nc>*** *July|NP* ***</nc>*** *to|TO stop|VB smoking|VBG|smoke and|CC it|PRP was|VBD|be an|DT* ***<cn>*** *incredible|JJ tool|NN* ***</cn>*** *to|TO keep|VB away|RB the|DT cravings|NNS|craving .|. </s#1> <s#2> However|RB ,|, it|PRP made|VBD|make me|PRP extremely|RB|extreme agitated|VBD|agitate .|. </s#2>*

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

Step #5 (NLP) – proper name and numeric concept categorization
<s#1> I|PRP went|VBD|go on|IN ***<pn cat=drug>*** *Zyban|NP </pn> in|IN* ***<nc cat=month>*** *July|NP </nc> to|TO stop|VB smoking|VBG|smoke and|CC it|PRP was|VBD|be an|DT <cn> incredible|JJ*

tool|NN </cn> to|TO keep|VB away|RB the|DT cravings|NNS|craving .|. </s#1> <s#2> However|RB ,|, it|PRP made|VBD|make me|PRP extremely|RB|extreme agitated|VBD|agitate .|. </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 automatically determined by <!metaMarker>.

Step #6 (NLP) – implicit metadata – issue, action, and outcome – generation

<s#1>

<action>
I|PRP went|VBD|go on|IN <pn cat=drug> Zyban|NP </pn> in|IN <nc cat=month> July|NP </nc>
</action>

<issue>
to|TO stop|VB smoking|VBG|smoke
</issue>

and|CC

<outcome>
it|PRP was|VBD|be an|DT <cn> incredible|JJ tool|NN </cn> to|TO keep|VB away|RB the|DT cravings|NNS|craving .|.
</outcome>

</s#1>

<s#2>

However|RB ,|,

<outcome>
it|PRP made|VBD|make me|PRP extremely|RB|extreme agitated|VBD|agitate .|.
</outcome>

</s#2>

This step assigns implicit metadata to each clause or phrase by categorizing each according to the predetermined schema of the communicative text discourse model. This categorization method is an adaptation of the sequential algorithm for training the text classifier [2]. The classifier utilized a training data set composed of a pre-coded set of examples. Each clause or phrase is represented as a feature vector consisting of NLP extracted explicit metadata from steps #1 to #5.

3 Text Classification

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. These 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 classing 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 classed. This process is illustrated in Figure 2.

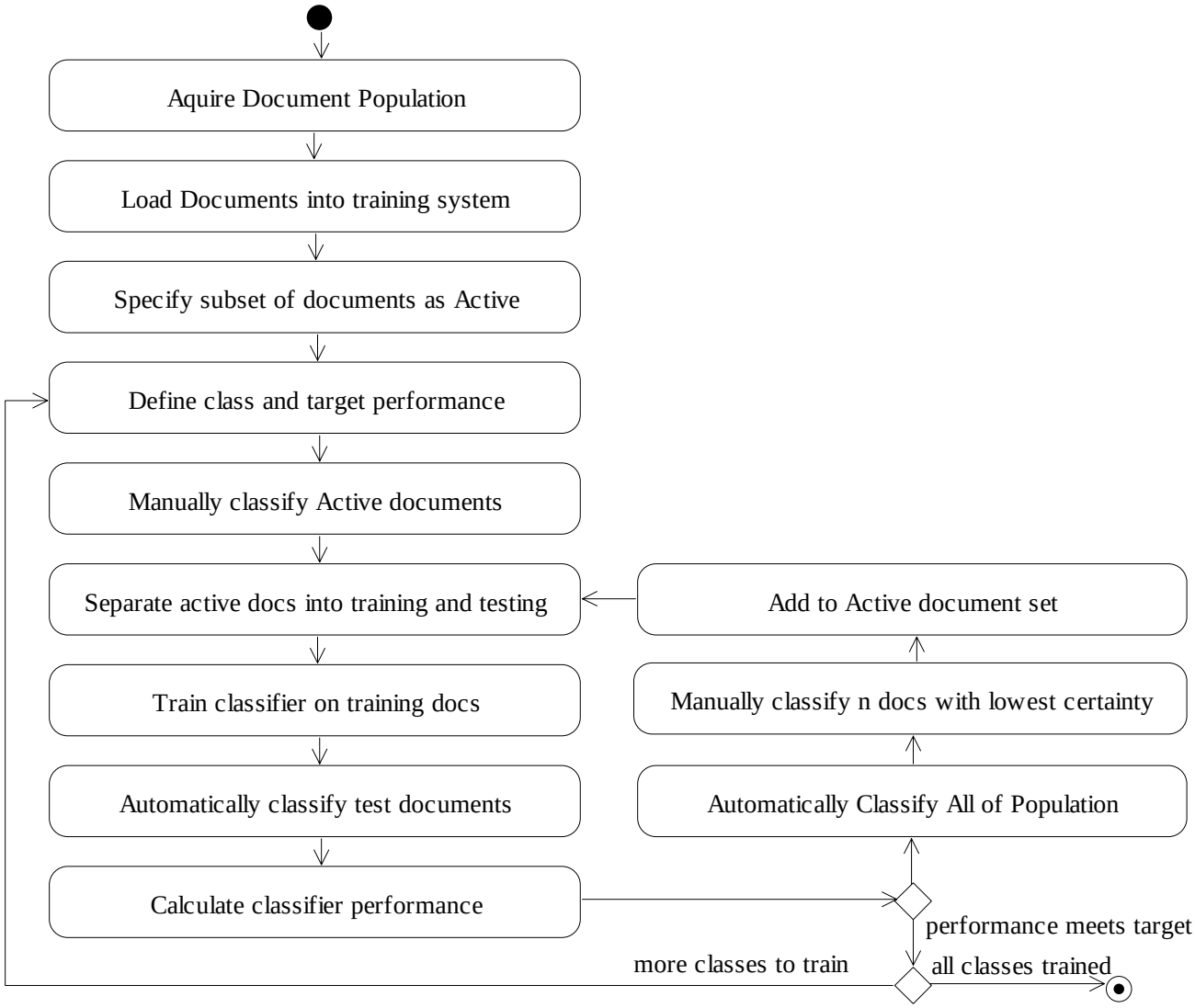

Figure 2: Text Classification Process

4 Experiments & 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 [1]. 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 are 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

Explicit metadata (such as proper names and numeric concepts) extraction rules were developed inductively by analyzing randomly selected training data from a collection of actual news stories, which were harvested from the web. There were about 4,000 news stories in the training data set. The text classifier used to generate the implicit metadata was trained by the same news stories after the appropriate implicit metadata was manually coded. In summary, the evaluation of the text classifier's effectiveness against the discussion group postings was done by the classifier, which was trained on the news stories.

The following steps were followed to measure the effectiveness of automatically extracting topic-oriented metadata from the discussion group postings.

- Test data was randomly selected and consisted of a pre-determined number of postings.

- 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.

There were about 2,000 postings in the testing data set. The preliminary experiment result for extracting the general subject type metadata using this previously unseen data are shown in the Table 1.

The precision of the text classifier test against the discussion group postings was significantly lower than the test against the news stories. Our previous test against the news stories produced an average of 80% precision in comparison to the 60% precision as shown in the Table 1. However, the recall figures based on the test against the discussion group postings were comparable to the recall scores based on the news stories.

Our preliminary observation generated a hypothesis that the precision differences are due to the conversational nature of the discussion group postings in comparison to the news stories and also that the postings are much less grammatical than the news stories.

However, this hypothesis needs to be studied and verified. Thus, we have begun to test further. Currently, the text classifier is being re-trained with the discussion group posting. We do not yet have the full test results. However, the preliminary results

showed that the precision scores have improved significantly.

Table 1: Preliminary Topic Classification Experiment Result

Class	**Precision**	**Recall**
Education	0.29	0.90
Engineering	0.61	0.90
LawCourts	0.40	0.94
Media	0.37	0.81
Music	0.63	0.95
Religion	0.29	0.90
Sports	0.58	0.94
Trade	0.30	0.92
Transportation	0.74	0.97
Travel	0.54	0.87
WeatherMeteor	0.68	1.00
AlternativeMed	0.77	1.00
AnthroSocio	0.62	0.93
Architecture	0.64	0.99
BiologicalSci	0.21	0.91
Chemistry	0.76	1.00
ClothingHome	0.87	1.00
EntertainmentInd	0.51	0.96
EquipTools	0.70	0.84
FoodDrink	0.89	0.95
Geology	0.75	1.00
History	0.84	1.00
Hobbies	0.65	1.00
Ling&Lang	0.81	1.00
MathStats	0.84	1.00
MedicalDental	0.66	0.94
PeopleSociety	0.48	1.00
Pets	0.74	0.95
PhilosophyEthics	0.93	0.97
PhysicsAstro	0.67	0.97
Psychology	0.66	0.98
CompSciTech	0.29	1.00
DrugsPharm	0.30	0.93
GovPolElec	0.44	0.81
Average	**0.60**	**0.95**

5 Conclusion

A combined NLP and ML approach to automate the discussion group participants' perception monitoring is introduced and its performance on a large number of discussion group postings is described. The described system is based on a general-purpose metadata generation process. In this paper we reported the test results of automatically generating topic-oriented metadata. This functionality is the first stage requirement of the described system as it is necessary to collect a particular set of the discussion group postings to generate a collective perception of the public about a particular topic.

The main goal of this application was 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.

When the described automatic metadata generation based perception monitoring system is fully functional one should be able to search on an issue or illness, the outcome (i.e. the type of side effect) or the action, i.e. the drug taken. Researchers and manufacturers might use the system to track complaints of certain types of medications. The general public can use the resultant data to assess their choice of treatment and regulators could use the information to ask for more clinical trials or pull a particularly harmful medicine from the shelf.

References

[1] Chincor, N. MUC-4 Evaluation Metrics. Proceedings of the Fourth Message Understanding Conference (MUC-4), McLean, VA, 1992

[2] Lewis, D. & Gale, W. A Sequential Algorithm for Training Text Classifier, SIGIR'94: Proceedings of Seventeenth Annual International ACM-SIGIR Conference on Research and Development

in Information Retrieval, Springer-Verlag, London, 1994.

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

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

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

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

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

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

[9] Santorini, B. Part-of-speech Tagging Guidelines for the Penn Treebank Project. Technical report, Department of Computer and Information Science, University of Pennsylvania, 1990.

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