ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

ML-Based Approach for Identifying Disease-Treatment Relations in Short Texts

P.Shirisha
Assistant Professor,
Department of Computer Science,
Malla Reddy Engineering College for Women,
Maisammaguda, Hyderabad.

K.Shirisha
Assistant Professor,
Department of Computer Science,
Malla Reddy Engineering College for Women,
Maisammaguda, Hyderabad.

Abstract:

The Machine Learning (ML) field has gained its momentum in almost any domain of research and just recently has become a reliable tool in the medical domain. Empirical domain of automatic learning is used in tasks such as medical decision support, protein-protein interaction, medical imaging, and extraction of medical knowledge. ML is envisioned as a tool by which computer-based systems can be integrated in the healthcare field in order to get a better and more efficient medical care.

A ML-based methodology for building an application that is capable of identifying and disseminating healthcare information. Due to advancements in medical domain automatic learning has gained popularity in the fields of medical decision support, complete health management and extraction of medical knowledge. The main objective of this work is to show what Natural Language Processing (NLP) and Machine Learning (ML) techniques used for representation of information and what classification algorithms are suitable for identifying and classifying relevant medical information in short texts.

This paper describes how ML and NLP can be used for extracting knowledge from published medical papers. It acknowledges the fact those tools capable of identifying reliable information in the medical domain stand as building blocks for a healthcare system that is up-to-date with the latest discoveries. Our research focus on the diseases and treatment information and the relation that exists between these two entities.

Index Terms:

Automatic Learning, Natural Language Processing, Machine Learning, Medical Decision Support, Healthcare, Classifiers.

INTRODUCTION:

This Work provides the foundation for development of technology framework that makes easy to find all the relevant information regarding treatment and diseases. The tool that is built with the techniques such as Natural Language Processing (NLP) and Machine Learning (ML) has capability to find all relevant short text information regarding diseases and treatments. This work presents various Machine Learning (ML) and information for classifying short texts and relation between diseases and treatments. According to ML technique the information are shown in short texts when identifying relations between two entities such as diseases and treatment. Thus there is improvement in solutions when using a pipeline of two tasks (Hierarchical way of approaching). It is better to identify and remove the sentence that does not contain information relevant to disease or treatments. The remaining sentences can be classified according to the interest. It will be very complex to identify the exact solution if everything is done in one step by classifying sentences based on interest and also including the sentences that do not provide relevant information.

Relation Extraction is a long standing research topic in Natural Language Processing. Medical information are stored in textual format among the biological data stored in Medline. Manually extracting useful information from large volume of database is a tedious work. Moreover HTML page displaying biological information contains medical information and typically unrelated materials such as navigation menus, forms, user comments, advertisement, feedback etc. The proposed work of this project extracts the useful disease related information with increased precision by using weighted bag of word representation [1] with a accuracy of 79% to 82%. The proposed approach supports in clinical decision making by providing physician with best available evidence of medical information.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The frequent use of electronic health records and information increase the need for text mining in order to improve the quality of result for the user query. This can result in two area of real time application[7] such as Text search engine targeted with Scientific document and Text Search engine targeted with technical document. In this project we choose text mining targeted with scientific document related to Medical treatment. Medline is chosen in this project to get biomedical information because it provides answers related to patient treatment and it's the database which is most widely used by the clinicians and research scholars in medical field. More importantly it is frequently updated and the contents are proved to be accurate compared to other medical websites providing information related to human disease, health, medicines, treatment etc. With the growing number of medical thesis, research papers, research articles, researchers are faced with the difficulty of reading a lot of research papers to gain knowledge in their field of interest. Search engines like Pub Med [8] reduces this constraint by retrieving the relevant document related to the user query.

Though the relevant document is retrieved, the web page displaying it may contain many non informative contents like advertisement, scroll bars, menus, citations, quick links, announcements, special credits, related searches, similar posts searched etc. This may be quite frustrating to the user when the user is in need of the information alone. In this project all the unrelated contents like advertisement etc mentioned in the above paragraph are removed and text mining is performed on the extracted document from which information or sentences related to user specified disease is extracted. From the extracted file symptoms, causes, treatment of the particular disease is filtered and displayed to the user. Thus the user gets the required information alone which saves his time and improves the quality of the result. This text mined document can be used in medical health care domain where a doctor can analyze various kinds of treatment that can be given to patient with particular medical disorder. The doctor can update the knowledge related to particular disease or its treatment methodology or the details of medicine that are in research for a particular disease. The doctor can gain idea about particular medicine that are effective for some patient but causes side effect to patient with some additional medical disorder. The patient can also use this extracted document to get clear understanding about a particular disease its symptoms, side effects, its medicines, its treatment methodologies.

Understanding the effect of a given intervention on the patient's health outcome is one of the key elements in providing optimal patient care. In the proposed approach a combination of structural natural language processing with machine learning method address the general and domain specific challenges of information extraction. Medical subheadings and subject heading may be used to infer relationship among medical concepts. The classification algorithm used in the proposed work exhibits effectiveness, efficiency, Online learning ability.

EXISTING SYSTEM:

The traditional healthcare system is also becoming one that hug the Internet and the electronic world. Electronic Health Records (EHR) is becoming the standard in the healthcare domain. Researches and studies show that the potential benefits of having an EHR system are:

Health information recording and clinical data repositories immediate access to patient diagnoses, allergies, and lab test results that enable better and time-efficient medical decisions; Medication management rapid access to information regarding potential adverse drug reactions, immunizations, supplies, etc; Decision support the ability to capture and use quality medical data for decisions in the workflow of healthcare; and Obtain treatments that are tailored to specific health needs—rapid access to information that is focused on certain topics.

Drawbacks:

1.In order to embrace the views that the EHR system has, we need better, faster, and more reliable access to information.

2.All research discoveries come and enter the repository at high rate, making the process of identifying and disseminating reliable information a very difficult task.

LITERATURE SURVEY:

[1] In Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger ,"Tackling The POOR Assumption Of Naïve Bayes Text Classifier" there were mentioned classification of text by using naïve bayes text classifier but use of navie bayes text classifier does not give precision 100% for output. Sometimes prediction of classifier may not be correct.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

[2]In T.Mouratis, S.Kotsiantis, "Increasing The Accuracy Of Discriminative Of Multinominal Bayesian Classifier In Text Classification", paper author introduced use of classifier that increased precision of output but problem in that work was at the time of classification it doesn't identify the verbs,nouns,adjectives properly so some time it nay give wrong value.

[3]In B.Rosario And M.A. Hearst,"Semantic Relation In Bioscience Text" where Hidden Markov models are used for entity recognition. This includes mapping biomedical information into structural representation. It involves converting natural language text into structural format. Their work uses machine learning for information extraction. The extraction of medical abstract is obtained through text classification. Semantic lexicons of words labeled with semantic classes so associations can be drawn between words which helps in extracting the necessary sentences related to the query. In this research paper the author used sentence co-occurrence and navie bayes algorithm to extract semantic relation like Gene-Protein from Medline abstract, the precision and recall of the result obtained are shown in the graph as their experimental results but due to use of only one navie bayes algorithm it do not get good precision of output, it doesn't used bag of words to find adjective, verbs while doing classification.

[4]In M.Craven, "Learning To Extract Relations From Medline" In their work the individual sentences are considered as instances that are to be processed by the navie bayes classifier. Here each instance is considered as positive training set. Alternative relation extraction are made through relational learning. Extraction of words from medline abstract has been done by using navie bayes,CNB algoritham and it also used bag of words during classification but not used natural language processing due to this performance of output degrades.

[5]In Oana Frunza.et.al, "A Machine Learning Approach For Identifying Disease-Treatment Relations In Short Texts" It involves automatic extraction of relation between medical concepts. A dictionary of medical terms is used for sentence classification. The sentences are automatically parsed using semantic parser. After applying semantic extraction a set of extraction, alteration, validation rules are applied to distinguish the actual semantic relation to be extracted but problem is that due to used of only one algorithm of machine learning navie bayes may not get good precision of output.

[6]In L. Hunter And K.B. Cohen, "Biomedical Language Processing: What's Beyond Pubmed" it involves Used of natural language processing for processing of biomedical words. in this work it takes the name of disease and give the solution which has been stored in database of that disease by parsing user statement using natural language processing but it does not do diagnosis of disease.

[7]In Jeff Pasternack, Don Roth "Extracting Article Text From Webb With Maximum Subsequence Segmentation" in involves to extract word according to occurrence of that word in article if no of word occur by no of time mentioned then I extract that word from the web here author used bag of word to remove verbs and adjective from the article but it doesn't use Natural language processing while extracting.

[8]In Abdur Rehman, Haroon.A.Babri, Mehreen saeed,"Feature Extraction Algorithm For Classification Of Text Document" It involves automatic extraction of relation between medical concepts. A dictionary of medical terms is used for sentence classification. The sentences are automatically parsed using semantic parser.it use 4 classification algorithm NB,CNB,Decission tree,Adaptive ,SVM etc wile extracting word but it doesn't give the information regarding diagnosis of disease.

[9]In Adrian Canedo-Rodriguez, Jung Hyoun Kim,etl.,"Efficient Text Extraction Aalgorithm Using Color Clustering For Language Translation In Mobile Phone" AdaBoost classifier is outperformed by other classifier. SVM classifier always functions well when the information matches with the training set. Probabilistic model always performs well on text classification task. Bag of word technique is simple in nature and in majority of the cases it is hard to outperform it. Pipelining of task is essential to obtain increased quality of result because majority class may overcome the underrepresented ones. By using pipelining there is a balance between relevant and irrelevant data and the classifier has better chance to distinguish relevant and non- relevant data but it don't used Gennia tagger which is special parser for medical words.

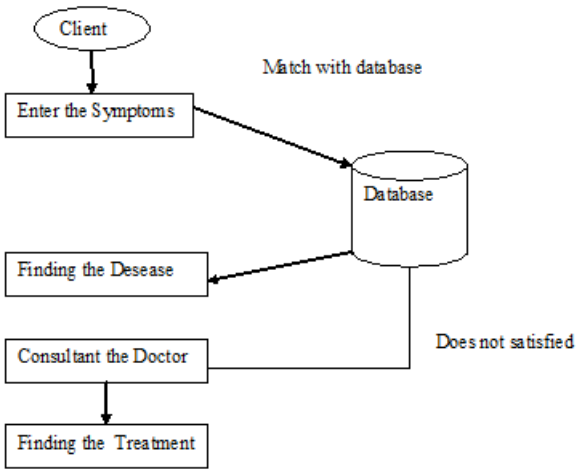
[10]In Oana Frunza, Diana Inkpen, and Thomas Tran, Member, IEEE "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts" it involves two task in pipelined manner for identifying and extracting the relationship between the given MEDLINE abstract.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Firs task involves finding most suitable model for prediction, the second task is to find good data representation. To achieve this two task various predictive algorithm and textual representation techniques are considered. A set of six classification algorithm namely decision based models, probabilistic models(Naïve Bayes, Complement Naïve Bayes), Adaptive learning, linear classifier namely support vector machine and a classifier that always predicts the majority class in training data are used. The advantages and limitations of all the six classification algorithm are discussed. Three representation technique namely Bag-Of-Word representation, NLP and Biomedical Concept representation and Medical concept representation are used to obtain the treatment relation from short text. Various experiments are conducted with the combination of the six classification algorithm and three representation techniques. The results are shown in bar chart form. As the result of the experiment it is concluded that bag-of-representation when combined with any of six classification algorithm produces better results but it does not give disease diagnosis as well information about particular disease by parsing statement.

ARCHITECTURE:

The server will extract the information from various articles related to those symptoms in the case of the work user can give their symptoms. Then it classifies that information based on the symptoms and then provides the cure, preventive measures and side effects for those symptoms.



The main task in this work is to extract healthcare information and the relation details. It focuses on diseases and treatment information, and the relation that exists between these two entities.

User interests are in line with the tendency of having a personalized medicine. It is not enough to read and know only about one study that states that a treatment is beneficial for a certain disease.

PROPOSED SYSTEM:

The two tasks used in this paper are the basis for the development of information technology framework. This framework helps to identify the medical related information from abstracts. The first task deals with extraction all information regarding diseases and treatments while the task deals with extraction of related information existing between disease and treatments. The framework developed with these tasks are used by healthcare providers, people who needs to take care of their health related problems and companies that build systematic views. The future product can be provided with browser plug-in and desktop application so that it helps the user to get all information related to diseases and treatments and also the relation between those entities.

It is also be useful to know more about latest discoveries related to medicine. The product can be developed and sold by companies that do research in medical care domain, Natural Language Processing (NLP), and Machine Learning (ML), and companies that develop tools like Microsoft Health Vault and Google Health. This product is valuable in e-commerce fields by showing the statistics that the information provided here are accurate and also provide all the recent discoveries related to health care. To make a product more popular it should be trust worthy so that people can buy it. It is the key factor foe any company to make product successful. When coming to health care products it should be more trust worthy since it is dealing with health related problems.

Companies that wish to sell health care framework need to develop tools that automatically extract the wealth of research. For example the information provided for diseases or treatments needs to be based on recent discoveries on health care field so that people can trust. The product quality also should be taken care so that it provides dynamic content for users. The first task deals with the identification of sentences from the Medline abstracts that provide the information about the diseases and treatments. In other words it also seems like scanning the sentences from Medline abstracts that contain relevant information which the user wants.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Natural Language Processing (NLP), and Machine Learning (ML) are used to extract accurate information or it can also say that it perfectly removes the unwanted information which are not related to disease or treatment. Natural Language Processing (NLP) and Machine Learning (ML) itself involve in extracting informative sentences. It is difficult task to identify the informative sentences in fields such as summarization and information extraction. The work and contribution value with this task is helpful in results and in settings for this task in healthcare field.

IMPLEMENTATION:

Tasks and Data Sets:

The two tasks that are undertaken in this paper provide the basis for the design of an information technology framework that is capable to identify and disseminate healthcare information. The first task identifies and extracts informative sentences on diseases and treatments topics, while the second one performs a finer grained classification of these sentences according to the semantic relations that exists between diseases and treatments. The problems addressed in this paper form the building blocks of a framework that can be used by healthcare providers (e.g., private clinics, hospitals, medical doctors, etc.), companies that build systematic reviews8 (hereafter, SR), or laypeople who want to be in charge of their health by reading the latest life science published articles related to their interests.

The final product can be envisioned as a browser plug-in or a desktop application that will automatically find and extract the latest medical discoveries related to disease-treatment relations and present them to the user.The product can be developed and sold by companies that do research in Healthcare Informatics, Natural Language Processing, and Machine Learning, and companies that develop tools like Microsoft Health Vault.

The value of the product from an e-commerce point of view stands in the fact that it can be used in marketing strategies to show that the information that is presented is trustful (Medline articles) and that the results are the latest discoveries. For any type of business, the trust and interest of customers are the key success factors. Consumers are looking to buy or use products that satisfy their needs and gain their trust and confidence. Healthcare products are probably the most sensitive to the trust and confidence of consumers.

Relationship	Definition and Example
Cure 810 (648, 162)	TREAT cures DIS *Intravenous immune globulin for recurrent spontaneous abortion*
Only DIS 616 (492, 124)	TREAT not mentioned *Social ties and susceptibility to the common cold*
Only TREAT 166 (132, 34)	DIS not mentioned *Fluticasome propionate is safe in recommended doses*
Prevent 63 (50, 13)	TREAT prevents the DIS *Statins for prevention of stroke*
Vague 36 (28, 8)	Very unclear relationship *Phenylbutazone and leukemia*
Side Effect 29 (24, 5)	DIS is a result of a TREAT *Malignant mesodermal mixed tumor of the uterus following irradiation*
NO Cure 4 (3, 1)	TREAT does not cure DIS *Evidence for double resistance to permethrin and malathion in head lice*
Total relevant: 1724 (1377, 347)	
Irrelevant 1771 (1416, 355)	Treat and DIS not present *Patients were followed up for 6 months*
Total: 3495 (2793, 702)	

The first task (task 1 or sentence selection) identifies sentences from Medline published abstracts that talk about diseases and treatments. We focus on three relations: Cure, Prevent, and Side Effect, a subset of the eight relations that the corpus is annotated with. We decided to focus on these three relations because these are most represented in the corpus while for the other five, very few examples are available. Table 1 presents the original data set, the one used by Rosario and Hearst [3], that we also use in our research. The numbers in parentheses represent the training and test set size. For example, for Cure relation, out of 810 sentences present in the data set, 648 are used for training and 162 for testing. The task of identifying the three semantic relations is addressed in two ways: Setting 1. Three models are built. Each model is focused on one relation and can distinguish sentences that contain the relation from sentences that do not. This setting is similar to a two-class classification task in which instances are labeled either with the relation in question (Positive label) or with non-relevant information (Negative label); Setting 2. One model is built, to distinguish the three relations in a three-class classification task where each sentence is labeled with one of the semantic relations.

Classification Algorithms and Data Representations:

In ML, as a field of empirical studies, the acquired expertise and knowledge from previous research guide the way of solving new tasks. The models should be reliable at identifying informative sentences and discriminating disease- treatment semantic relations. The research experiments need to be guided such that high performance is obtained.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The experimental settings are directed such that they are adapted to the domain of study (medical knowledge) and to the type of data we deal with (short texts or sentences), allowing for the methods to bring improved performance. There are at least two challenges that can be encountered while working with ML techniques. One is to find the most suitable model for prediction. The ML field offers a suite of predictive models (algorithms) that can be used and deployed. The task of finding the suitable one relies heavily on empirical studies and knowledge expertise. The second one is to find a good data representation and to do feature engineering because features strongly influence the performance of the models. Identifying the right and sufficient features to represent the data for the predictive models, especially when the source of information is not large, as it is the case of sentences, is a crucial aspect that needs to be taken into consideration.

These challenges are addressed by trying various predictive algorithms, and by using various textual representation techniques that we consider suitable for the task. As classification algorithms, we use a set of six representative models: decision-based models (Decision trees), probabilistic models (Naı¨ve Bayes (NB) and Complement Naı¨ve Bayes (CNB), which is adapted for text with imbalanced class distribution), adaptive learning (Ada- Boost), a linear classifier (support vector machine (SVM) with polynomial kernel), and a classifier that always predicts the majority class in the training data (used as a baseline). We decided to use these classifiers because they are representative for the learning algorithms in the literature and were shown to work well on both short and long texts. Decision trees are decision-based models similar to the rule-based models that are used in handcrafted systems, and are suitable for short texts.

Semantic Relation	F-measure Task1	F-measure Task2	F-measure Pipeline
Cure	90.72%	98.55%	89.40%
Prevent	90.72%	100%	90.72%
SideEffect	90.72%	88.89%	80.64%

Probabilistic models, especially the ones based on the Naı¨ve Bayes theory, are the state of the art in text classification and in almost any automatic text classification task. Adaptive learning algorithms are the ones that focus on hard-to-learn concepts, usually underrepresented in the data, a characteristic that appears in our short texts and imbalanced data sets.

SVM-based models are acknowledged state-of-the-art classification techniques on text. All classifiers are part of a tool called Weka.9 One can imagine the steps of processing the data (in our case textual information—sentences) for ML algorithms as the steps required to obtain a database table that contains as many columns as the number of features selected to represent the data, and as many rows as the number of data points from the collection (sentences in our case).

Bag Of Words:

The bag-of-words (BOW) representation is commonly used for text classification tasks. It is a representation in which features are chosen among the words that are present in the training data. Selection techniques are used in order to identify the most suitable words as features. After the feature space is identified, each training and test instance is mapped to this feature representation by giving values to each feature for a certain instance. Two most common feature value representations for BOW representation are: binary feature values—the value of a feature can be either 0 or 1, where 1 represents the fact that the feature is present in the instance and 0 otherwise; or frequency feature values—the value of the feature is the number of times it appears in an instance, or 0 if it did not appear. Because we deal with short texts with an average of 20 words per sentence, the difference between a binary value representation and a frequency value representation is not large. In our case, we chose a frequency value representation. This has the advantage that if a feature appears more than once in a sentence, this means that it is important and the frequency value representation will capture this.

CONCLUSION:

This approach is very useful for everyone as it gives information only of the area of interest. The interests are in line with the tendency of having a personalized medicine that has one in which each patient has its medical care tailored to its needs. This study is related to a particular field but the future scope of the paper lies in the fact that this can be extended to the information on the web. The proposed system used the top concept candidate for each identified phrase in an abstract as a feature. Experimental result shows that the technique used in the proposed work minimizes the time and the work load of the doctors in analyzing information about certain disease and treatment in order to make decision about patient monitoring and treatment.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

This text mined document can be used in medical health care domain where a doctor can analyze various kinds of treatment that can be given to patient with particular medical disorder. The doctor can update the knowledge related to particular disease or its treatment methodology or the details of medicine that are in research for a particular disease. The doctor can gain idea about particular medicine that are effective for some patient but causes side effect to patient with some additional medical disorder. The patient can also use this extracted document to get clear understanding about a particular disease its symptoms, side effects, its medicines, its treatment methodologies. This paper also present healthcare diagnosis treatment & prevention of disease, illness, injury in human.

FUTURE WORK:

We also consider as potential future work ways in which the framework's capabilities can be used in a commercial recommender system. It extracts diseases and treatments given in and identifies the semantic relations between them.

REFERENCE:

[1] R. Bunescu and R. Mooney, "A Shortest Path Dependency Kernel for Relation Extraction," Proc. Conf. Human Language Technologyand Empirical Methods in Natural Language Processing (HLT/EMNLP), pp. 724-731, 2005.

[2] R. Bunescu, R. Mooney, Y. Weiss, B. Scho¨ lkopf, and J. Platt,"Subsequence Kernels for Relation Extraction," Advances in NeuralInformation Processing Systems, vol. 18,pp. 171-178, 2006.

[3] A.M. Cohen and W.R. Hersh, and R.T. Bhupatiraju, "FeatureGeneration, Feature Selection, Classifiers, and Conceptual Driftfor Biomedical Document Triage," Proc. 13thText Retrieval Conf.(TREC), 2004.

[4] M. Craven, "Learning to Extract Relations from Medline," Proc.Assoc. for theAdvancement of Artificial Intelligence, 1999.[5] I. Donaldson et al., "PreBIND and Textomy: Mining theBiomedical Literature for Protein-Protein Interactions UsingaSupport Vector Machine," BMC Bioinformatics, vol. 4, 2003.

[6] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky,"GENIES: A Natural Language Processing System for theExtraction of Molecular Pathways from JournalArticles," Bioinformatics,vol. 17, pp. S74-S82, 2001.

[7] O. Frunza and D. Inkpen, "Textual Information in Predicting Functional Properties of the Genes," Proc. Workshop Current Trend sin Biomedical Natural Language Processing (BioNLP) in conjunction with Assoc. for Computational Linguistics (ACL '08), 2008.

[8] R. Gaizauskas, G. Demetriou, P.J. Artymiuk, and P. Willett,"Protein Structures and Information Extraction from BiologicalTexts: The PASTA System," Bioinformatics, vol. 19, no. 1, pp. 135-143, 2003.

[9] C. Giuliano, L. Alberto, and R. Lorenza, "Exploiting Shallow Linguistic Information for Relation Extraction from BiomedicalLiterature," Proc. 11th Conf. European Chapter of the Assoc. forComputational Linguistics, 2006.

[10] J. Ginsberg, H. Mohebbi Matthew, S.P. Rajan, B. Lynnette, S.S.Mark, and L. Brilliant, "Detecting Influenza Epidemics UsingSearch Engine Query Data," Nature, vol. 457, pp. 1012-1014, Feb.2009.

[11] M. Goadrich, L. Oliphant, and J. Shavlik, "Learning Ensembles ofFirst-Order Clauses for Recall-Precision Curves: A Case Study inBiomedical Information Extraction," Proc. 14th Int'l Conf. InductiveLogic Programming, 2004.