Identifying Semantic Relations Between Disease And Treatment Using Machine Learning Approach

Manikandan. S
Assistant Professor/Department of IT
EGSPEC, Nagapattinam, Tamilnadu, India

Manikanda Kumaran. K
Assistant Professor/Department of IT
EGSPEC, Nagapattinam, Tamilnadu, India

Abstract

The Machine Learning field has gained its momentum in almost any domain of research and just recently has become a reliable tool in the medical domain. The empirical domain of automatic learning is used in tasks such as medical decision support, medical imaging, protein-protein interaction, extraction of medical knowledge, and for overall patient management care. ML is envisioned as a tool by which computer based systems can be integrated in the healthcare field in order to get a better, more efficient medical care. This project describes a ML-based methodology for building an application that is capable of identifying and disseminating healthcare information. It extracts sentences from published medical papers that mention diseases and treatments, and identifies semantic relations that exist between diseases and treatments. The evaluation results for these tasks show that the proposed methodology obtains reliable outcomes that could be integrated in an application to be used in the medical care domain.

Index: Natural Language processing, Medical imaging, Machine learning.

1. Introduction

This paper aimed at designing and examining various representation techniques in combination with various learning methods to identify and extract biomedical relations from literature. The contributions that we bring with the work stand in the fact that we present an extensive study of various ML algorithms and textual representations for classifying short medical texts and identifying semantic relations between two medical entities: diseases and treatments [1]. From an ML point of view, the consent in short texts when identifying semantic relations between diseases and treatments. It is better to identify and eliminate first the sentences that do not contain relevant information, and then classify the rest of the sentences by the relations of interest, instead of doing everything in one step by classifying sentences into one of the relations of interest plus the extra class of uninformative sentences.

In this paper to build an application that is capable of identifying and disseminating healthcare information. To integrate the computer based systems into healthcare fields. It extracts sentences from published medical papers that mention diseases and treatments, and identifies semantic relations that exist between diseases and treatments.

1.1 A Shortest Path Dependency Kernel for Relation Extraction

A novel approach to relation extraction, based on the observation that the information required to assert a relationship between two named entities in the same sentence is typically captured by the shortest path between the two entities in the dependency graph. Experiments on extracting top-level relations from the ACE newspaper corpus show that the new shortest path dependency kernel outperforms a recent approach based on dependency tree kernels one of the key tasks in natural language processing is that of Information Extraction, which is traditionally divided into three sub problems: co reference resolution, named entity recognition, and relation extraction. Consequently, IE corpora are typically annotated with information corresponding to these subtasks, facilitating the development of systems that target only one or a subset of the three problems [2].

Local and non-local dependencies are equally important for finding relations. In this project a purpose of extracting both types of dependencies using a CCG parser, however another approach is to recover deep dependencies from syntactic parses was performed. This may have the advantage of preserving the quality of local dependencies while completing the representation with non-local dependencies. In this project the focus is exclusively on extracting relations between predefined types of entities in the ACE corpus. Reliably extracting relations between entities in natural-language documents is still a difficult, unsolved problem. a new kernel for relation extraction based on the shortest-path between the two relation entities in the dependency graph. Comparative experiments on extracting top-level relations from the ACE corpus show significant improvements over a recent dependency tree kernel. The method assumes that the named entities are known. A natural extension is to automatically extract both the entities and their relationships.

1.2 Paper Organization

The rest of the paper is organized as follows. Section 2 introduces system analysis. Section 3 introduces system implementation algorithm and specifies the algorithms. Section 4 describes case study of medical image processing and algorithm. Finally, conclude the paper in Section 5.

2. System Analysis

The traditional healthcare system is also becoming one that embraces the Internet and the electronic world. Electronic Health Records are becoming the standard in the healthcare domain. MEDLINE contains bibliographic information and abstracts from more than 4,000 biomedical journals. Much of the important, late-breaking bioscience information is found only in textual form, and so methods are needed to automatically extract semantic entities and the relations between them. In the medical domain, the richest and most used Source of information is Medline, a database of extensive life science published articles [3].

Document classification is a common problem in biomedicine. First, the number of true positives in both the training and test collection was known to be small, between 6 and 7%. Second, the utility function chosen as the metric of record was heavily weighted to reward recall and not precision. This was based on an analysis of the current working procedures and an approximation of how they currently value false negative and false positive classification.

Drawbacks

1. The current existing system does not support immediate sentence selection and relational classification.
2. Difficulty in extracting particular medical journal from a pool of bibliographic journal.
3. Lack automatic information mapping system.

The Main concepts that are proposed were to automatically identifying sentences published in medical abstracts (Medline) as containing or not information about diseases and treatments, and automatically identifying semantic relations that exist between diseases and treatments, as expressed in these texts. The second task is focused on three semantic relations such as Cure, Prevent, and Side Effect. The tasks that are addressed here are the foundation of an information technology framework that identifies and disseminates healthcare information. The objective for this project is to show what Natural Language Processing and Machine Learning techniques; representations of information and classification algorithms are suitable to use for identifying and classifying relevant medical information in short texts.

3. System Implementation

3.1 Medical Management

Natural language processing for biomedical text currently focuses mostly on entity and relation extraction. In the medical field, Medline is the main source of publications. It currently contains more than 12 million citations and is growing fast. Natural language processing is a set of techniques that can help facilitate analysis, retrieval, and integration of textual and electronic information. During the last few years, NLP has also become important in bioinformatics and agree with **Maojo et al** [5] that both disciplines can and should learn from each other. The aim of the medical management scheme was the rapid access to information regarding potential adverse drug reactions, immunizations, supplies. In the medical domain, the richest and most used source of information is Medline, a database of extensive life science published articles. All research discoveries come and enter the repository at high rate making the process of identifying and disseminating reliable information a very difficult task. In the medical management the user and admin authentication was performed. The user has to register first to avoid unnecessary data abruption[fig.1]. The Central Agent Admin has to retrieve the medical abstracts and has to use the training data for further processing.

Input: Medical Abstract, Training Data
Output: Data for text classification.

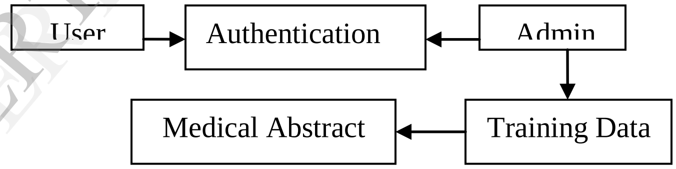

Fig. 1 Overview of Medical Management

3.2 Sentence Selection

Information Extraction is an important task in natural language processing, with many practical applications. It involves the analysis of text documents, with the aim of identifying particular types of entities and relations among them. Reliably extracting relations between entities in natural-language documents is still a difficult, unsolved problem.

Information extraction may be defined as the task of automatically extracting instances of specified classes or relations from text. Automated methods for information extraction have several valuable applications including populating knowledge bases and databases, summarizing collections of documents, and identifying significant but unknown relationships among objects. Since constructing information extraction systems manually has proven to be expensive, there has been much recent interest in using machine learning methods to learn information extraction models from labeled training data [fig.2]. The primary function is to identify sentences from Medline published abstracts that talk about diseases and treatments.

The task is similar to a scan of sentences contained in the abstract of an article in order to present to the user-only sentences that are identified as containing relevant information. Extracting informative sentences is a task by itself in the NLP and ML community. Research fields like summarization and information extraction are disciplines where the identification of informative text is a crucial task. The contributions and research value that are brought with this task stand in the usefulness of the results and the insights about the experimental settings for the task in the medical domain. For the first task, the data sets are annotated with the following information: a label indicating that the sentence is informative, i.e., containing disease-treatment information, or a label indicating that the sentence is not informative [6].

Input: Medical Abstracted Trained Data
Output: Sentence (Text) identication-Informative.

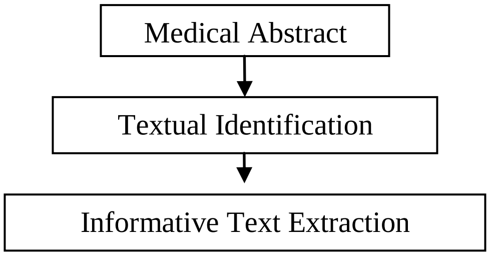

Fig. 2 Overview of Sentence Selection

3.3 Classification Algorithm: SRLS Algorithm [Sampling for Regularized Least Squares Classification]

The algorithm assigns a univariate “score” or “importance” to every feature. It then randomly samples a small number of features, and solves the classification problem induced on those feature.

A theorem which provides worst-case guarantees on the generalization power of the resultant classification function f’ with respect to that of f obtained by using all the features. To the best of the knowledge, this is the first feature selection method with such guarantees. This project provide additive-error approximation guarantees for any query document and relative-error approximation guarantees for query documents that satisfy a somewhat stronger but reasonable condition with respect to the training document corpus. Thus, the proof of the main **quality-of-approximation theorem** provides an analytical basis for commonly held intuition about when such feature selection algorithms should and should not be expected to perform well. It have a corpus of d training documents, each of which is described by **n * d** features. The main goal is to choose a small number r of features, where d. **r * n,** such that, by using only those r features, to obtain good classification quality, both in theory and in practice, when compared to using the full set of n features. In particular, like to solve exactly or approximately a RLSC problem of the form to get a vector to classify successfully a new document according to a classification function of the form.

The algorithm takes as input the **n × d** term-document matrix A, a vector **y * Rd** of document labels where **sign(yj)** labels the class of document A(j) (where A(j) denotes the **jth** column of the matrix A and A(i) denotes the ith row of A), and a query document q * Rn. It also takes as input a regularization parameter λ * R+, a probability distribution **{pi}n i=1** over the features, and a positive integer r. The algorithm first randomly samples roughly r features according to the input probability distribution. Let ˜ **A** be the matrix whose rows consist of the chosen feature vectors, rescaled appropriately, and let **˜q** be the vector consisting of the corresponding elements of the input query document q, rescaled in the same manner.

An important aspect of the algorithm is the probability distribution {pi} n i=1 input to the algorithm. One could perform random sampling with respect to any probability distribution. On the other hand, more intelligent sampling can lead to improved classification performance, both theoretically and empirically.

Input: Extracted informative text
Output: Classifier Classes.

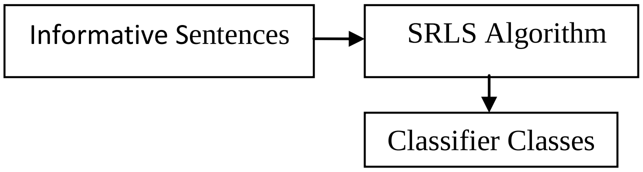

Fig. 3 Overview of Classification Algorithm

3.4 Rule Based Inductive Classifier

Rule learning algorithms, have become a successful strategy for classifier induction. Rule based classifiers provide the desirable property of being interpretable and, thus, easily modifiable based on the user's a priori knowledge. A novel method is used for the automatic induction of rule-based text classifiers. Rule induction is based on a greedy optimization heuristics whereby a set of high-quality rules are generated for the category being learned.

The rule based text classifier implements on the automatic text classification with the help of SRLS sampling algorithm, they automatically classify the sentences based upon certain keywords relating the disease and treatment. Automatic text classification means, automatic assignment of documents to a predefined set of categories.

The classifier classes were first included in the repository manager's document repository which must be included for the preprocessor. Then the special categorizer classifies it into text classified results. Once a classifier for category c has been constructed, its capability to take the right categorization decision is tested by applying it to the documents of the test set and then comparing the resulting classification to the ideal one.

Input: Extracted informative text
Output: Classified Text for analysis.

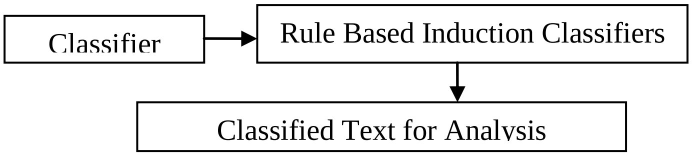
Classifier
Rule Based Induction Classifiers
Classified Text for Analysis

Fig. 4 Overview of Rule Based Inductive Classifier

3.5 Textual Relation Identification

The task of relation extraction or relation identification is previously tackled in the medical literature, but with a focus on biomedical tasks. Usually, the data sets used in biomedical specific tasks use short texts, often sentences. This is the case of the first two related works mentioned above. The tasks often entail identification of relations between entities that co-occur in the same sentence.

It has a deeper semantic dimension and it is focused on identifying disease-treatment relations in the sentences already selected as being informative (e.g., task 1 is applied first). The focus is on three relations: Cure, Prevent, and Side Effect, a subset of the eight relations that the corpus is annotated with. The focus is on these three relations because these are most represented in the corpus while for the other five, very few examples are available.

Input: Classified text
Output: Medical measures (Cure, Prevent, Side Effects).

4. Biomedical Information Extraction

Many domains in the field of Inductive Logic Programming involve highly unbalanced data. The research has focused on Information Extraction, a task that typically involves many more negative examples than positive examples. IE is the process of finding facts in unstructured text, such as biomedical journals, and putting those facts in an organized system. In particular, focus is on learning to recognize instances of the protein-localization relationship in Medline abstracts. The view is that the problem as a machine-learning task: given positive and negative extractions from a training corpus of abstracts, learn a logical theory that performs well on a held-aside testing set. A common way to measure performance in these domains is to use precision and recall instead of simply using accuracy.

Gleaner is proposed which is a randomized search method which collects good clauses from a broad spectrum of points along the recall dimension in recall-precision curves and employs at least N of these M clause thresholding method to combine the selected clauses. Gleaner is compared to ensembles of standard Aleph theories and found that Gleaner produces comparable test set results in a Fraction of the training time needed for ensembles. Domains suitable for Inductive Logic Programming can be roughly divided into two main groups. In one group, there are tasks in which each example has some inherent relational structure. One classic example of this domain is the trains’ dataset, where the goal is to discriminate between two types of trains, and the trains themselves are relational objects, having varying length and types of objects carried by each car. Multi-Slot Information Extraction is a an appealing challenge task for ILP, due to its large amount of examples and background knowledge, as well as the substantial skew of examples. A method called Gleaner was developed, which gathers a wide spectrum of clauses and combines them within bins based on recall using atleast N of these M clauses thresholding method.

5. Conclusion

The conclusions of project suggest that domain-specific knowledge improves the results. The representation techniques influence the results of the ML algorithms, but more informative representations are the ones that consistently obtain the best results. The project shows that the best results are obtained when the classifier is not overwhelmed by sentences that are not related to the task. Thus the project is highly validated by the usage of SRLS Algorithm and rule based inductive classifier.

To extend the experimental methodology when the first setting is applied for the second task, to use additional sources of information as representation techniques, and to focus more on ways to integrate the research discoveries in a framework to be deployed to consumers. In addition to more methodological settings in which to find the potential value of other types of representations, a focus on source data that comes from the web.

6. References

[1]. R. Bunescu and R. Mooney, "A Shortest Path Dependency Kernel for Relation Extraction," Proc. Conf. Human Language Technology and 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 Neural Information Processing Systems, vol. 18, pp. 171-178, 2006.

[3]. A.M. Cohen and W.R. Hersh, and R.T. Bhupatiraju, "Feature Generation, Feature Selection, Classifiers, and Conceptual Drift for Biomedical Document Triage," Proc. 13th Text Retrieval Conf. (TREC), 2004

[4]. M. Craven, "Learning to Extract Relations from Medline," Proc.Assoc. For the Advancement of Artificial Intelligence, 1999.

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

[6]. J. Li, Z. Zhang, X. Li, and H. Chen, "Kernel-Based Learning for Biomedical Relation Extraction," J. Am. Soc. Information Science and Technology, vol. 59, no. 5, pp. 756-769, 2008.

[7]. B. Rosarioand M.A. Hearst, "Semantic Relations in Bioscience Text," Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics, vol. 430, 2004.

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

[9]. http://www.healthcare-informatics.com.

Authors Profile

S. Manikandan received the **B.Tech.** degree in Information Technology from the EGS Pillay Engineering College, Nagapattinam, Anna University, Chennai, India, in 2010 and received **M.E.** Computer Science and Engineering in Annamalai University, Annamalai Nagar India, in 2012. Currently he is working Assistant Professor in EGS Pillay Engineering College, Nagapattinam, India. His research interest includes Soft computing, Data mining and Warehousing, Fuzzy Logic, Pervasive Computing.

K. Manikanda Kumaran received the **B.Tech.** degree in Information Technology from the EGS Pillay Engineering College, Nagapattinam, Anna University, Chennai, India, in 2010 and received **M.E.** Computer Science and Engineering in Annamalai University, Annamalai Nagar India, in 2012. Currently he is working Assistant Professor in EGS Pillay Engineering College, Nagapattinam, India. His research interest includes Soft computing, Data mining and Warehousing, wireless sensor networks.