An Environment for Data Analysis in Biomedical Domain: Information Extraction for Decision Support Systems

Pablo F. Matos[1], Leonardo O. Lombardi[2], Thiago A.S. Pardo[3], Cristina D.A. Ciferri[3], Marina T.P. Vieira[2], and Ricardo R. Ciferri[1]

[1] Department of Computer Science, Federal University of São Carlos - São Carlos/SP, Brazil
[2] Faculty of Mathematical and Nature Sciences,
Methodist University of Piracicaba - Piracicaba/SP, Brazil
[3] Department of Computer Science, University of São Paulo - São Carlos/SP, Brazil
pablo_matos@dc.ufscar.br, lolombardi@unimep.br,
{taspardo,cdac}@icmc.usp.br, mtvieira@unimep.br,
ricardo@dc.ufscar.br

Abstract. This paper addresses the problem of extracting and processing relevant information from unstructured electronic documents of the biomedical domain. The documents are full scientific papers. This problem imposes several challenges, such as identifying text passages that contain relevant information, collecting the relevant information pieces, populating a database and a data warehouse, and mining these data. For this purpose, this paper proposes the IEDSS-Bio, an environment for Information Extraction and Decision Support System in Biomedical domain. In a case study, experiments with machine learning for identifying relevant text passages (disease and treatment effects, and patients number information on Sickle Cell Anemia papers) showed that the best results (95.9% *accuracy*) were obtained with a statistical method and the use of preprocessing techniques to resample the examples and to eliminate noise.

Keywords: Classification, information extraction, knowledge discovery, biomedical domain.

1 Introduction

In the biomedical domain there are a lot of electronic documents that report experiments involving patients who have some kind of disease, describing the treatment adopted, the number of patients enrolled in the treatment, which symptoms and risk factors are associated with the disease, and if the treatment has interfered positively or negatively in the patient's health. The experiments are reported in several magazines and journals, e.g., *American Journal of Hematology*, *Blood*, and *Haematologica.* Researchers and doctors are not able to process this huge number of documents to extract key information related to some issues of interest.

These documents are in unstructured format, that is, in plain textual form. It is necessary to transform this information from unstructured to structured format in order to

N. García-Pedrajas et al. (Eds.): IEA/AIE 2010, Part I, LNAI 6096, pp. 306–316, 2010.

submit it to an automatic knowledge discovery process. For this purpose, an environment/framework called **I**nformation **E**xtraction and **D**ecision **S**upport **S**ystem in **Bio**-medical domain (IEDSS-Bio) is proposed in this paper. This environment is under development and aims at supporting the expert in making decisions, by extracting relevant information from biomedical documents, storing the information in a data warehouse, and mining interesting knowledge from it.

After presenting the general environment architecture and its data flow, this paper focuses mainly on the adopted information extraction process, more specifically, on the task of identifying text passages/sentences that contain the relevant information. A case study on identifying sentences about disease and treatment effects and patients number from Sickle Cell Anemia papers is reported, showing that a high *accuracy* (95.9%) is achieved with a statistical machine learning method and the use of pre-processing techniques to resample the examples and to eliminate noise.

This paper is organized as follows. Section 2 presents the theoretical foundation about text and data mining, and information extraction. Section 3 reviews the main related work. Section 4 presents the IEDSS-Bio environment and some examples of the mined knowledge. Section 5 describes the information extraction module. Section 6 discusses the experiments and Section 7 concludes the paper.

2 Theoretical Foundation

Text mining [1] refers to the process of extracting useful information from documents in unstructured format by identifying knowledge and exploiting patterns. Texts are converted into structured format to use data mining techniques.

Data Mining is the application of specific algorithms for extracting patterns from data [2]. One of the most popular data mining tasks is the discovery of association rules, which aims at finding items that frequently occur together in the data. An association rule is an implication of the form $A => B$, where A and B are item sets [3]. The implication means that databases tuples (transactions) satisfying A are likely to satisfy B. *Support* and *confidence* measures are used to indicate relevant rules [3].

Traditionally, data mining algorithms are applied to data that are gathered in a single table. The data gathering process from multiple tables through joints or aggregation may cause loss of meaning or information and may have a high computational cost [4]. Multi-relational data mining methods search for patterns that involve multiple tables (relations) from a relational database, maintaining the data separately in the data mining process. In the biomedical database it is adequate the use of multi-relational data mining due to the properties of the data.

Regarding information extraction, Cohen and Hunter [5] present two approaches: the rule-based approach and the machine learning approach. The first one uses some kind of knowledge; the second one uses classifiers to separate sentences or documents. Krauthammer and Nenadic [6] and Ananiadou and McNaught [7] present a third approach: the dictionary-based approach, which uses information from a dictionary to assist in the identification of terms or entities in the text. These approaches are the three predominant ones for knowledge extraction in the biomedical domain.

The dictionary-based approach has the advantage of storing information related to a particular area and makes possible the identification of terms such as names of gene and protein. Some problems of this approach are the limited number of names in the

dictionary, the change of names, which generates a low *recall*, and the short names, which generate false positives and decrease the *precision* [8].

The rule-based approach has some disadvantages: to delay the construction of systems significantly, to reduce the adaptability of rules to another system, and to remove terms that do not match the predefined patterns. However, generally it has performed better than other approaches [7].

The advantages of using the machine learning approach are the domain-independence and high quality prediction. The main problems related to the machine learning algorithms are the need for large amount of representative training data [7] and for (possible) retraining with the advent of new data.

3 Related Work

In the biomedical literature, there are studies that extract information from abstracts or full papers about gene or protein mainly, using a combination of the three approaches previously discussed. Most of these studies extract information from MEDLINE abstracts [9, 10, 11, 12]. The works that extract information from full text papers have different goals, namely: to extract information [13], to populate a database [14], or to highlight the sentences in accordance with a user query [15]. Some of the above works use *Part-Of-Speech tagging* [9, 10, 13, 16], while some do not [11, 12, 14, 15].

Among the studies, [16] produced the highest *precision* (72.5%) and *recall* (50.7%), using a combination of machine learning, dictionary-based and rule-based approaches. Using the same approaches to extract information from abstracts, higher values were obtained, respectively, 85.7% and 66.7% [13]. This difference shows the particularity in extracting information from full papers. According to Cohen and Hersh [17], AbGene [13] is the most successful rule-based approach for the recognition of gene and protein in biomedical texts.

4 An Environment for Data Analysis

The IEDSS-Bio environment is shown in Fig. 1. It is divided into two major components as shown in the dotted rectangles:

- *Conversion and Extraction component*: aims at converting different formats and extracting relevant information from scientific documents to store it in a biomedical database (**1**);
- *Data Analysis component*: aims at identifying patterns from the biomedical domain by applying data warehouse and data mining techniques (**2**).

From the papers available in PDF format, the *Convert module* handles PDF files and converts them into the semi-structured XML format. The information contained in the XML document is organized in hierarchical levels such as *section* » *page* » *paragraph* » *sentence* to make the extraction processing easier. Next, the *Information Extraction module* processes the information from the XML document to extract the most relevant information from the sentences. Fig. 1 shows that the expert may interfere in the *Converter module*, verifying the quality of the generated document and suggesting changes. In Section 5 the *Information Extraction* (IE) *module* is explained in details.

After the extraction of relevant information from the XML document, the extracted information is stored in the biomedical database by the *Persistence Layer module*. The *Data Extraction and Integration* (DEI) *module* is responsible for extracting data from the biomedical database through the *Persistence Layer* and for populating the *Data Warehouse* (DW).

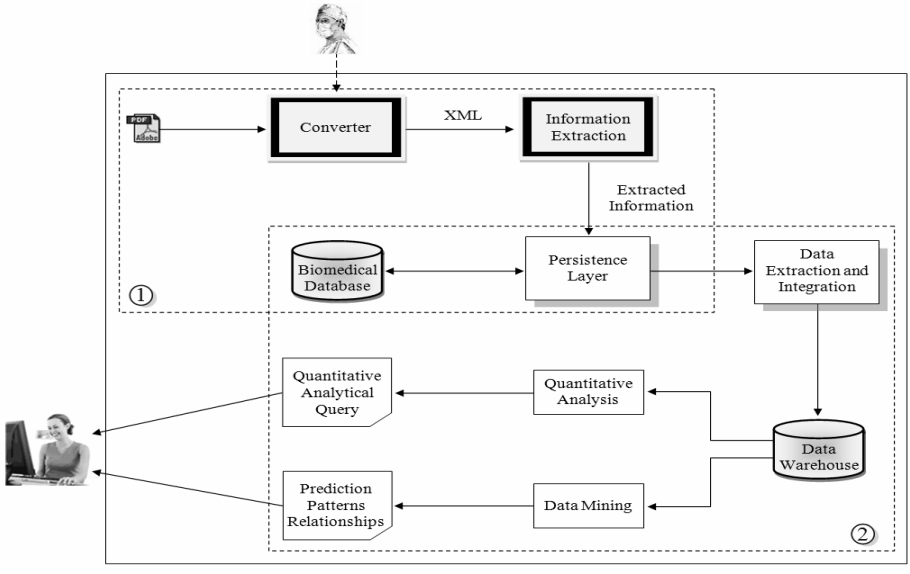

Fig. 1. Environment for data analysis

The data stored in the DW may be used in two kinds of processes in order to support data analysis. In the first one, the DW may be consulted for obtaining quantitative analysis. For instance, it is possible to answer the question "*How many patients had clinical improvement and were treated with the hydroxyurea drug?*" In the second one, the DW may be used by data mining algorithms aiming at identifying interesting patterns in the data. For instance, it is possible to find out that "*A significant amount of patients under treatment with the hydroxyurea drug tend to have marrow depression*".

The biomedical database of Fig. 1 has been populated with treatments, positive and negative effects, and number of patients extracted from experiments reported in the papers. Fig. 2 shows parts of some of the tables that compose the database. The joint analysis of this information may lead to the discovery of useful information for decision-making. However, the following features in the data need to be adequately addressed, requiring a new approach for the data mining algorithm to be used:

1. The number of patients involved is not explicitly reported in many circumstances (which had undergone certain treatments, which had certain effects or certain symptoms), requiring some inference on the data; only after this process, the data may be stored in the database;

2. The information about treatments and (negative or positive) effects was stored in two ways in the database:
   - in a table that lists treatment and effect caused by it, when the paper states clearly that information (table (a) in Fig. 2);
   - in separate tables, if the association between cause and effect is not mentioned in the paper (tables (b), (c) and (d) in Fig. 2).

Even though this information is located in different tables and do not maintain explicit links between each other, frequent co-occurrences in the data may indicate that relationships exist between them. An example would be a significant number of experiments reporting the use of a specific treatment and the occurrence of certain effect. Therefore, it is important to analyze such information. The multi-relational data mining technique is suitable for this purpose.

Fig. 2 illustrates a subset of stored data. Considering $hydroxyurea = T2$, $lower\_annual\_rates\_of\_crises = Pe2$ and $marrow\_depression = Ne3$, the following rule could be found by jointly analyzing the tables (b), (c), and (d):

Rule 1: hydroxyurea => lower_annual_rates_of_crises, marrow_depression [support = 10%; confidence = 80%]

where values of *support* and *confidence* are just for illustration and do not refer to an actual generated rule.

Paper #	Treatment	Positive Effect	Patients Number
P1	T1	Pe1	30
P1	T2	Pe2	20
P2	T1	Pe2	10
P2	T1	Pe3	50
P3	T1	Pe1	25
P3	T3	Pe1	10
P3	T3	Pe3	60

(a)

Paper #	Treatment	Patients Number
P1	T3	58
P4	T1	46
P4	T2	82
P4	T3	16
P5	T2	64
P5	T1	107

(b)

Paper #	Positive Effect	Patients Number
P1	Pe3	52
P4	Pe1	89
P4	Pe2	70
P5	Pe1	45
P5	Pe2	35
P6	Pe2	20

(c)

Paper #	Negative Effect	Patients Number
P1	Ne1	25
P4	Ne2	30
P4	Ne3	22
P5	Ne3	12
P5	Ne4	4
P6	Ne2	7

(d)

Fig. 2. Treatment_Positive_Effect (a), Treatment (b), Positive_Effect (c), and Negative_Effect (d) tables

5 Information Extraction Module

The information extraction of biomedical domain uses a combination of three approaches:

1. Machine learning: used for sentence classification;
2. Rule-based: formal method to specify text patterns, used to extract information from sentences classified in the previous step;
3. Dictionary-based: a manually built dictionary (containing terms from the scientific papers) stores information related to the biomedical domain in order to assist the construction of rules, used to increase the *precision* and *recall* in the identification of such information.

Fig. 3 shows the adopted process for extracting information from the biomedical domain. The sentences from each paper processed by the *IE module* are classified using machine learning techniques. Each classified sentence is analyzed through regular expressions (rules) to identify relevant information. The dictionary of terms from the biomedical domain is used to assist the construction of the regular expressions. An expert may validate the information extracted (filter) before it is stored in the database. The expert may also evaluate the classified sentences and add new terms to the dictionary.

Formally, consider $S = \{s_1, s_2, \ldots, s_n\}$ the set of sentences of training and $C = \{c_1, c_2, \ldots, c_n\}$ the set of predefined classes. The classification aims at identifying a function f that maps each sentence s_i to one of the predefined classes c_j: $f(s_i) = c_j$, where $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

For each class c_j a set of regular expressions re_θ is defined in order to obtain relevant information. The symbol ϴ denotes one of the following concepts: the number of patients involved in the experiment under different circumstances (total patients, patients who have had positive or negative effects, patients that have completed or not the treatment), the treatment adopted, the positive or negative effects and the symptoms of the patient.

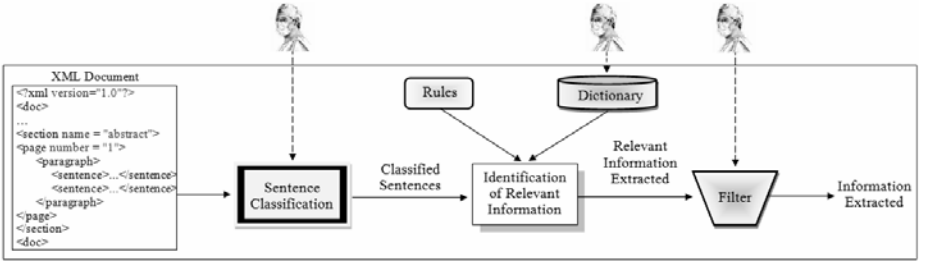

Fig. 3. Information extraction module

The sentence classification model is built from the documents (papers) suggested by experts. The model consists of structuring the sentences in an attribute-value matrix whose rows represent the sentences and the columns the attributes (n-grams). Each sentence is associated to a class. Each cell in the matrix indicates the presence/frequency or absence of the corresponding attribute in a sentence. The data may undergo stopwords removal, stemming, and attribute selection processes.

The *IE module* identifies the relevant information of the classified sentences through regular expressions and the dictionary. The following example shows one possible regular expression:

$$re_{negative\_effect\_patients} = (\langle \text{negative effect} \rangle)(.*)([0\text{–}9]+)\text{percent}(.*)(\langle \text{individual} \rangle)$$

The ⟨negative effect⟩ and ⟨individual⟩ are terms that are stored in the dictionary. The following sentences are examples represented by this regular expression:

"Respiratory failure was documented in 13 percent of patients."

"Neurologic events occurred in 11 percent of patients."

"Neurologic complications developed in 22 percent of the adults in our study."

These sentences express the percentage of patients who suffered from some negative effects. Although the information is different, the sentence structures are similar.

Regular expressions are used to cover different ways of expressing the same information. Some challenges were faced, although regular expressions provide resources that allow dealing with different writing styles:

1. In some cases, it is necessary to interpret the sentence to understand it, which is beyond the scope of regular expressions. One example is shown in the following sentence: *"Fewer patients assigned to hydroxyurea had chest syndrome (25 vs. 51, $P < 0.001$), and fewer underwent transfusions (48 vs. 73, $P = 0.001$)."*

   The author compares the results of two treatments to which the patients were submitted. He cites the number of patients who had chest syndrome and those who underwent transfusions for each treatment. Given the way he reports these numbers, the information requires interpretation of other information mentioned in the paper to identify the treatment that led to each result.
2. In other cases, additional treatment is necessary to obtain the desired information. For instance, in order to estimate the number of patients who suffered from respiratory failure in the following sentence (13%) it is necessary to identify all patients who participated in the experiment, which was mentioned somewhere else in the text: *"Respiratory failure was documented in 13 percent of patients."*

6 Evaluation: A Case Study

Experiments were carried out in the biomedical domain, more specifically on papers about Sickle Cell Anemia [18], which is a genetic and hereditary disease considered as a public health problem. These experiments are specifically on sentence classification from medical scientific papers, dealing with positive and negative effects and the number of patients enrolled in treatments. The goal is to validate the classification task, which is an important step in the information extraction process.

Three questions are intended to be answered here: (1) How do human beings manually perform the sentence classification? (2) Is it feasible to automate the sentence classification task? (3) What kind of classification algorithm performs better in this classification?

To answer the *first* question, it is necessary to know the results obtained by humans in the sentence classification. For this end, the Kappa measure [19] was used to compute the agreement among humans. The achieved results are shown in Section 6.1. After that, it is possible to answer the *second* question by comparing the obtained Kappa value with a nominal scale agreement. To know the answer for the *third* question, six classical machine learning algorithms were chosen. These algorithms are from different paradigms: Support Vector Machine (SVM) and Naïve Bayes (NB) are statistical methods; ID3, J48, Prism, and OneR are symbolic methods – the first two are algorithms for decision tree induction and the last two are algorithms for rules induction. All classification experiments were performed with Weka data mining environment [20]. Default values for the learning parameters were used as they appear in Weka. The Mover classification system [21] is considered as a baseline method. It is a classical tool for text structure classification and is a NB approach to the problem. The results are shown in Section 6.2.

6.1 Annotation Agreement

The Kappa measure for the evaluation of annotation agreement among humans was calculated in order to know the difficulty of the sentence classification. The idea is that the more humans agree the better the task is defined and, therefore, the more it may be automated. Humans were divided into two groups: naïve subjects (i.e., that are not experts on the topic) and experts. Table 1 shows the results for a sample of 50 sentences. One may see that there were 3 naïve subjects and 3 experts. In the last row of the table, the annotation agreement was collected for all humans, experts or not. We computed the agreement for positive and negative effect classes, for other possible classes (under the class named as "other"), and for the 3 classes together. An interesting result is that the naïve subjects group has gotten the overall agreement value higher (0.71) than the experts (0.63). According to the Landis and Koch scale [22], the obtained Kappa for all the classes is in the range of substantial agreement (0.61 to 0.81). We may conclude that the classification task is well defined. Therefore, this result answers the *first* and *second* questions mentioned previously, i.e., for this task it is possible to automatically perform the classification.

Table 1. Annotator agreement on 50 sentences

Annotator	*Positive Effect*	*Negative Effect*	*Other*	*All the classes*
3 experts	0.77	0.63	0.52	**0.63**
3 naïve subjects	0.80	0.72	0.60	**0.71**
experts + naïve subjects	0.75	0.66	0.55	**0.65**

6.2 Experiments with Sentence Classification

In this section, the experiments on sentence classification are reported, focusing on the information of interest: effect and patients. Two samples were selected for each one: Sample279 and Sample600 for effect sentences; and Sample204 and Sample659 for patients sentences. The number that identifies each sample also indicates the number of examples/sentences that the corresponding sample contains. The effect sentences were divided into three classes: *Positive Effect*, *Negative Effect* and *Other*. The patients sentences were separated in five classes: *Negative Effect Concluded Patients*, *Negative Effect NonConcluded Patients*, *Positive Effect Concluded Patients*, *Total Patients* and *Other*. *Concluded* and *NonConcluded* refer to the patients that concluded and did not conclude the experiments, respectively. The percentage distribution of the classes for the information of interest may be seen in Fig. 4.

(a)

Sample	*Positive Effect*	*Negative Effect*	*Other*
Sample279	17.20%	20.08%	62.72%
Sample600	15.67%	19.00%	65.33%

(b)

Sample	*NE Concluded Patients*	*NE NonConcluded Patients*	*PE Concluded Patients*	*Total Patients*	*Other*
Sample204	4.90%	3.92%	7.85%	3.92%	79.41%
Sample659	6.07%	2.73%	2.73%	1.06%	87.41%

Fig. 4. Effect (a) and Patient (b) distribution of classes for each sample

The experiments were conducted as follows. First of all, the sentences (for every class) were cleaned by removing commas, parenthesis, etc. Then the attribute-value matrix was constructed using minimum frequency two for selecting the attribute, i.e., attributes whose values happened only once were not considered. The attributes were 1 to 3-grams, and the values were: 1, for the case the n-gram occurs in the sentence, or 0 otherwise. Neither stopwords removal nor stemming were considered.

After that, the six previous machine learning algorithms were used in the experiments. Six kinds of preprocessing were used for each algorithm, generating 36 possible combinations. The preprocessing possibilities were: *No Filter* (NF), *Randomize* (RD)*, Remove Misclassification* (RM), *Resample* (RS), RM followed by RS, and RS followed by RM. NF does not use any kind of filter; RD is used to random the examples; RM is used to eliminate noise; RS is an oversampling method used to balance the examples. The partitioning method 10-fold cross-validation was used to generate the results. It trains the classification model on 9 folds and tests it on the remaining one. The *accuracy* measure (number of correctly classified sentences/number of sentences) was computed.

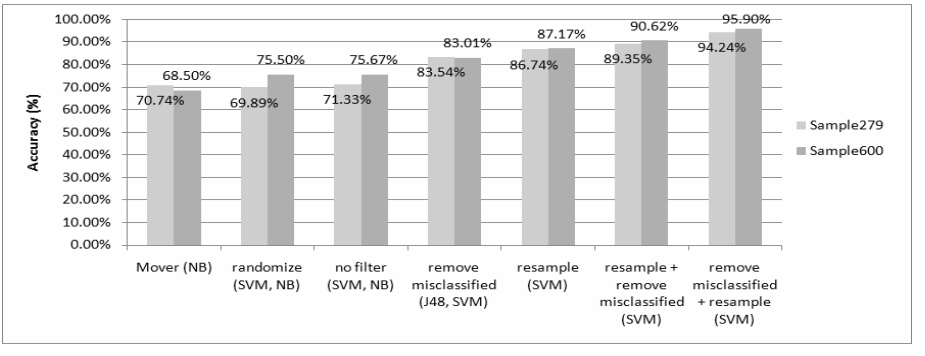

Fig. 5. Better results (*accuracy* measure) obtained in effect class

The results reported in Fig. 5 show that in general the SVM algorithm performed better than the other algorithms. The best results were produced with the use of RM and RS filters, i.e., by removing the mislabeled sentences and resampling the examples in order to establish a uniform rate between the classes. The results obtained from patients class were similar to the effect class and are not shown in this paper. The better results obtained by SVM answer the *third* question in this paper.

Since regular expressions select the sentence parts of interest after sentence classification, it is best suited that the classifier recovers a big number of sentences, even if they are misclassified (high *recall*, but probably low *precision*). A classifier that selects correctly most of the sentences (high *precision*) may recover fewer sentences, and this may injure the information extraction process.

7 Conclusion and Future Work

The environment proposed in this paper – Information Extraction and Decision Support System in Biomedical domain – aims at being a general framework for mining

relevant information in the area. Our first experiments on sentence classification (which is a step of the whole process) showed very good results (95.9% *accuracy*) for papers about Sickle Cell Anemia. Therefore, the task of sentence classification in this area is well defined and possible to be automated.

As future work, we plan to investigate the identification of treatment and symptoms information in the text, as well as proceed to the extraction of the relevant sentence pieces for populating our databases (using some of the techniques discussed in this paper). We also envision to investigate the use of parallel processing to optimize the more time-consuming tasks, e.g., the application of data mining algorithms and the analytical query processing. Other biomedical areas may also benefit from our text mining approach. This also remains for future work.

Acknowledgments. The authors acknowledge the support of the following Brazilian research agencies: CAPES, FAPESP, CNPq, and FINEP.

References

1. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York (2007)
2. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. AI Magazine 17(3), 37–54 (1996)
3. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2006)
4. Džeroski, S.: Multi-Relational Data Mining: An Introduction. ACM SIGKDD Explorations Newsletter 5(1), 1–16 (2003)
5. Cohen, K.B., Hunter, L.: Getting Started in Text Mining. PLoS Computational Biology 4(1), 1–3 (2008)
6. Krauthammer, M., Nenadic, G.: Term Identification in the Biomedical Literature. Journal of Biomedical Informatics 37(6), 512–526 (2004)
7. Ananiadou, S., McNaught, J. (eds.): Text Mining for Biology and Biomedicine. Artech House, Norwood (2006)
8. Tsuruoka, Y., Tsujii, J.: Improving the Performance of Dictionary-Based Approaches in Protein Name Recognition. Journal of Biomedical Informatics 37(6), 461–470 (2004)
9. Chun, H.-W., Tsuruoka, Y., Kim, J.-D., Shiba, R., Nagata, N., Hishiki, T., Tsujii, J.: Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine Learning. In: 11th PSB, Hawaii, pp. 4–15 (2006)
10. Mika, S., Rost, B.: NLProt: Extracting Protein Names and Sequences from Papers. Nucleic Acids Research 32(Suppl. 2), 634–637 (2004)
11. Seki, K., Mostafa, J.: A Hybrid Approach to Protein Name Identification in Biomedical Texts. Information Processing & Management 41(4), 723–743 (2005)
12. Hanisch, D., Fundel, K., Mevissen, H., Zimmer, R., Fluck, J.: Prominer: Rule-Based Protein and Gene Entity Recognition. BMC Bioinf. 6(Suppl. 1), S14 (2005)
13. Tanabe, L., Wilbur, W.J.: Tagging Gene and Protein Names in Biomedical Text. Bioinformatics 18(8), 1124–1132 (2002)
14. Bremer, E.G., Natarajan, J., Zhang, Y., DeSesa, C., Hack, C.J., Dubitzky, W.: Text Mining of Full Text Articles and Creation of a Knowledge Base for Analysis of Microarray Data. In: López, J.A., Benfenati, E., Dubitzky, W. (eds.) KELSI 2004. LNCS (LNAI), vol. 3303, pp. 84–95. Springer, Heidelberg (2004)

15. Garten, Y., Altman, R.: Pharmspresso: A Text Mining Tool for Extraction of Pharmacogenomic Concepts and Relationships from Full Text. BMC Bioinf. 10(Suppl. 2), S6 (2009)
16. Tanabe, L., Wilbur, W.J.: Tagging Gene and Protein Names in Full Text Articles. In: Workshop on NLP in the Biomedical Domain, pp. 9–13. ACL, Phildadelphia (2002)
17. Cohen, A.M., Hersh, W.R.: A Survey of Current Work in Biomedical Text Mining. Briefings in Bioinformatics 6(1), 57–71 (2005)
18. Pinto, A.C.S., Matos, P.F., Perlin, C.B., Andrade, C.G., Carosia, A.E.O., Lombardi, L.O., Ciferri, R.R., Pardo, T.A.S., Ciferri, C.D.A., Vieira, M.T.P.: Technical Report Sickle Cell Anemia. Technical Report, Federal University of São Carlos (2009), http://sca.dc.ufscar.br/download/files/report.sca.pdf
19. Fleiss, J.L.: Measuring Nominal Scale Agreement among Many Raters. Psychological Bulletin 76(5), 378–382 (1971)
20. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, 2nd edn. Morgan Kaufmann, San Francisco (2005)
21. Anthony, L., Lashkia, G.V.: Mover: A Machine Learning Tool to Assist in the Reading and Writing of Technical Papers. IEEE Trans. Prof. Comm. 46(3), 185–193 (2003)
22. Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Biometrics 33(1), 159–174 (1977)