Identifying Bioentity Recognition Errors of Rule-Based Text-Mining Systems

Francisco M. Couto, Tiago Grego, Hugo P. Bastos, Catia Pesquita
Faculty of Sciences, University of Lisbon
fcouto@di.fc.ul.pt

Rafael Torres, Pablo Sánchez, Leandro Pascual, Christian Blaschke
Bioalma SL, Tres Cantos (Madrid), Spain

Abstract

An important research topic in Bioinformatics involves the exploration of vast amounts of biological and biomedical scientific literature (BioLiterature). Over the last few decades, text-mining systems have exploited this BioLiterature to reduce the time spent by researchers in its analysis. However, state-of-the-art approaches are still far from reaching performance levels acceptable by curators, and below the performance obtained in other domains, such as personal name recognition or news text.

To achieve high levels of performance, it is essential that text mining tools effectively recognize bioentities present in BioLiterature. This paper presents FiBRE (Filtering Bioentity Recognition Errors), a system for automatically filtering misannotations generated by rule-based systems that automatically recognize bioentities in BioLiterature. FiBRE aims at using different sets of automatically generated annotations to identify the main features that characterize an annotation of being of a certain type. These features are then used to filter misannotations using a confidence threshold.

The assessment of FiBRE was performed on a set of more than 17,000 documents, previously annotated by Text Detective, a state-of-the-art rule-based name bioentity recognition system. Curators evaluated the gene annotations given by Text Detective that FiBRE classified as non-gene annotations, and we found that FiBRE was able to filter with a precision above 92% more than 600 misannotations, requiring minimal human effort, which demonstrates the effectiveness of FiBRE in a realistic scenario.

1 Introduction

Analyzing BioLiterature is a painful and hard task, even to an expert, given the large number of articles being published and the complexity of their content. However, text-mining tools applied to BioLiterature are also still far from reaching performance levels comparable to the obtained in other areas. For instance, text mining tools already achieve both precision and recall higher than 90% in personal name recognition on news text [18, 7, 10]. BioLiterature is also more complex than news text: news text is written in a way that the general public can understand its message, which is not the case in BioLiterature that has a much smaller and specific audience.

The main challenge in BioLiterature analysis is the lack of a standard nomenclature for describing biologic concepts and entities. We can often find different terms referring to the same biological concept or entity (synonyms), or the same term meaning different biological concepts or entities (homonyms). Also, genes whose name is a common English word are frequent, which makes it difficult to recognize biological entities in the text [11]. These issues make it difficult to correctly recognize entities and concepts mentioned in a text and are at the root of the poor performance exhibited by text-mining tools in BioLiterature analysis [6, 13]. Moreover, many state-of-the-art text-mining tools that recognize entities in BioLiterature are rule-based, and consequently are affected by the significant number of exceptions that rules derived from BioLiterature must encompass. A serious bottleneck of these approaches is the difficulty to devise from a subpart of the text a set of rules incorporating all the possible exceptions to them. Therefore, the application of these rules in a slightly different domain normally lead to a significant number of errors.

To address this problem we developed FiBRE (Filtering Bioentity Recognition Errors), a system for automatically filtering the errors made by rule-based systems that annotate BioLiterature[4]. By annotation we mean the piece of text where a bioentity was recognized, not the functional annotation. For each type of bioentity annotated, (e.g. gene, protein, chemical compound, drug, disease, symptom) FiBRE identifies the main features that characterize each annotation type. For example, consider that we have a set of

gene annotations and a set of disease annotations, FiBRE will most probably find that the gene annotations almost never have the word *disease* in the surroundings of the annotation, unlike disease annotations that frequently have the word *disease* in the surroundings of the annotation. FiBRE will use this kind of information to identify putative misannotations, i.e. the gene annotations with the word *disease* in the surroundings of the annotation.

The remainder of this paper is organized as follows. Section 2 describes the state-of-the-art of Text Mining applied to BioLiterature. Section 3 describes FiBRE in detail. Section 4 presents the experimental evaluation of FiBRE using the annotations made by a rule-based named entity recognition system. Finally, Section 5 expresses our main conclusions.

2 State-of-the-art

Most state-of-the-art text-mining systems use a rule-based or a case-based (statistical or machine-learning) approach for retrieving information from the text [5]. The manual analysis of text requires less expertise in the case-based approach than in the rule-based approach. In the rule-based approach, the expert has to identify not only the expected output, but also how the relevant information is expressed. This expertise can, however, be used by rule-based systems to achieve higher precision by selecting the most reliable rules and patterns.

Many surveys report the performances of text-mining tools that are run in different corpora (collection of documents) to execute different tasks [9, 1, 6, 14, 3]. On the other hand, recent challenging evaluations compared the performance of different approaches in solving the same tasks using the same corpus [18], [7], [10],[8]. The results achieved in these competitions show that text-mining systems are still far from reaching desirable performance levels. Thus, novel techniques are required to reinforce and further improve the quality and impact of Text Mining of BioLiterature.

BioAlma is developing a state-of-the-art system named Text Detective, which is capable of annotating a wide range of biological entities, such as genes, proteins, chemical compounds, drugs, diseases, symptoms and generic biomedical terms [16]. Here is an example of the output of an annotation returned by Text Detective, describing the occurrence of the presumed gene *21-channel digital EEG* found in the abstract with PubMed identifier 10599856:

```
>10599856 Medline gene Homo sapiens
21-channel digital EEG 21-channel digital EEG 3:59-80 PA
```

The first line consists of the PubMed identifier, the type of annotation and the organism name. The second line contains the name and abbreviation of the gene, the place where the gene was found (sentence number : start character - end character), and labels describing the rule used to find this annotation.

Text Detective is a rule-based system, which means that the process of identifying the entities on the text is based on a predefined set of rules that are manually managed. For the gene identification process, the system achieves an average of 80% precision, i.e. the system correctly annotates 80% of the genes, and fails for the 20% remnant. Curators do not normally consider this level of performance as satisfactory, thus tools that could improve the performance of Text Detective are much required. The identification of rules requires more effort from the curators than the evaluation of a limited set of cases. However, a single rule can express knowledge beyond that contained in a large set of cases. None of the knowledge representation techniques subsumes the other: the knowledge enclosed in a rule is normally not fully expressed by a finite set of cases, and it is difficult to identify a set of rules encoding all the knowledge expressed by a set of cases. Therefore, FiBRE intends to get the benefits of both approaches by using the case-based approach to validate the results of rule-based systems, such as Text Detective. FiBRE uses the annotations of the rule-based systems to automatically create the training sets, i.e. it is based on weakly supervised machine-learning approaches that were recently tested for identifying gene mentions in text [17, 2].

3 FiBRE

FiBRE is an add-on tool for rule-based named entity recognition systems that produce annotations at least of two different types, e.g. gene and non-gene entities. FiBRE can also be applied to case-based systems, but we think that it would be much less effective than using rule-based systems since in the bottom line we would be applying the same technique twice. Having different categories is a prerequisite of any case-based approach, since the training set should at least have positive and negative cases so that a model can be created.

Input: FiBRE receives a set of annotations containing at least two types of annotations that the named entity recognition system found.
In our experiment, the annotations given by Text Detective were split in two sets: one with the gene annotations and the other with the remaining non-gene annotations (chemical compounds, drugs, diseases and symptoms).

Output: The output is the list of putative misannotations, i.e. the annotations that were systematically irregular for all training/test splits.

PubMed id	Sentence	Score	Curation
15479369	Two of these were developed for children (the Haemo-QoL and the CHO-KLAT), and two for adults (the Hemofilia-QoL and the **Hemolatin-QoL**).	0.99	ok
15505396	This paper summarizes the published experience as well as results of the 3rd International Workshop on Glutaryl-CoA Dehydrogenase Deficiencyheld in October 2003 in Heidelberg, Germany, on the topic treatment of patients with glutaryl-CoA dehydrogenase (**GCDH**) deficiency.	0.98	not ok
15655003	In trial 3, heifers in **IDO 3** (n = 71) were again treated as in IDO 1.	0.97	ok
15587756	Assessment of **QT** interval duration and dispersion in athlete's heart.	0.97	ok
10599856	13 healthy subjects (28.5 3.8 years) were recorded with a **21-channel digital EEG** during a stroboscopic alternative motion paradigm implying illusionary motion with ambiguous direction.	0.94	ok

Table 1. Putative misannotations identified by FiBRE. The text highlighted represents the annotation found by Text Detective in that sentence. For each annotation, FiBRE returns a confidence score of it being a misannotation. The last column shows the curator decision, *ok* means that FiBRE was correct and *not ok* means that the annotation is correct and therefore FiBRE was not. The last row corresponds to the annotation presented in Section 2

.

In our experiment, the output is the given gene annotations that FiBRE classified as non-genes. Each annotation that FiBRE returns is accompanied by a classification score provided by the classification method.

Figure 1 represents an outline of the main steps of FiBRE: creation of the training and test sets by selecting annotations from both types; creation of the model using a statistical classification method based on the training sets of both categories; classification of the test set annotations using the model; filtering of the annotations that were systematically irregular, i.e. that FiBRE classified as being of a different type in all different training/test splits.

To use a machine-learning approach, FiBRE has to represent each annotation identified by Text Detective by a set of features. Thus, FiBRE represents each annotation mainly by three different types of features:

Context: the words that appear before and after the annotation in the same sentence;

Annotation: the words present in the annotation;

Part-of: the prefixes and suffixes of words that appear in the annotation.

Note that the annotation is not necessarily the name of the bioentity recognized. Text Detective may have recognized the bioentity by another expression. FiBRE ignored stop words, such as *in* or *on.* for representing the annotation. FiBRE also removed the commoner morphological and inflectional endings from words to avoid creating distinct features for words with a similar meaning [15].

In the actual version FiBRE selects as context features at most 8 other words of context, 4 after and 4 before the annotation, provided that the sentence contains enough words before/after the annotation. In previous versions of FiBRE we found that including more than 8 words as context features did not improve the results at all. The value for each context feature is proportional to its distance to the annotation. FiBRE starts by assigning a value of 10 to the closest word (after or before) and decreases 3 each time it moves far away from the annotation, i.e. FiBRE assigns values from the following set: $10, 7, 4, 1$. If the same word appears before and after the annotation, the feature based on this word will accumulate both values. Besides the context word itself, FiBRE creates another feature with the same value that discriminates if the word was found after or before the annotation. FiBRE represents this feature by delimiting the word by the $>$ or $<$ character if the word occurs before or after the annotation, respectively.

FiBRE selects as annotation features the words that compose the annotation and assigns to them a value of 10. If a context word is present in the annotation, the features based on this word will accumulate both values. Besides each annotation word itself, FiBRE creates another feature with the same value that discriminates that the word is present in the annotation. FiBRE represents this feature by delimiting the word by the $|$ character.

For each annotation word, FiBRE creates a feature for the prefixes and suffixes of the word. By prefix and suffix we consider the first and last n characters of the word, respectively. The features are obtained by varying n from 2 to 5, i.e. for each word we have at most 4 prefixes and 4 suffixes depending on the size of the word. FiBRE assigns a value of 5 for each prefix and a value of 10 for suffixes. FiBRE represents prefixes and suffixes features by adding the character $*$ to the end of each prefix and to the beginning of each suffix.

Figure 2 presents the FiBRE representation of the annotation created by Text Detective presented in the last row of Table 1. The annotation is represented by eight context words and one annotation word and the features devised from them.

After characterizing each annotation by a set of features, FiBRE uses them in a machine-learning process performed

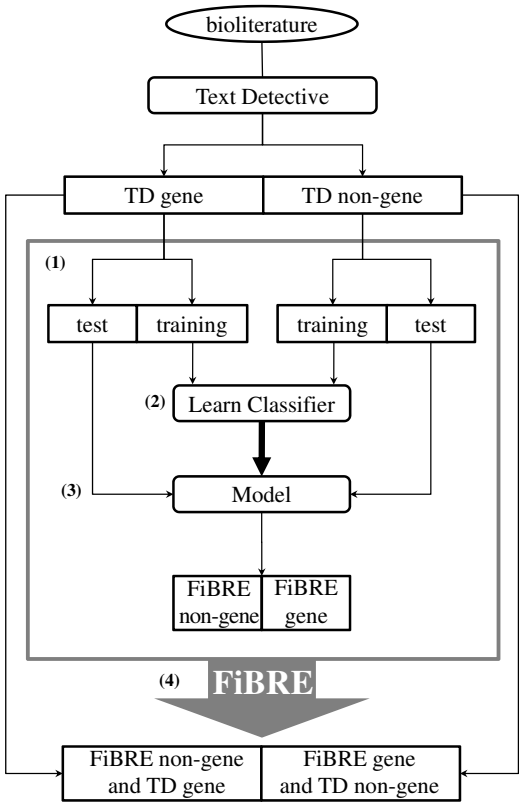

Figure 1. FiBRE receives gene and non-gene annotations from Text Detective and returns the annotations that FiBRE systematically classified as irregular, i.e. the gene annotations that were classified as being non-gene annotations and vice versa.

by a statistical classification method, to find the features that characterize each type of annotation. The model generated by the machine-learning process is then used by FiBRE to classify all the annotations in the test set disregarding their original type. This will produce two categories of annotations: *regular* that were classified according to their original type; and *irregular* that were classified as being of a different type.

FiBRE executes the machine-learning process multiple times with different training/test set splits in order to use the same annotation sometimes for training other times for testing. Finally, FiBRE classifies annotations that are systematically irregular for all training/test splits as putative misannotations. Notice that besides being a case-based system, FiBRE does not require any human intervention to create the training sets, since they are automatically generated from the set of annotations produced by the rule-based system.

```
10599856#21-channel_digital_eeg#003#59:80#gene_pos gene_pos
>healthi> 1 >subject> 4 >year> 7 >record> 10
 healthi  1  subject  4  year  7  record  10
<stroboscopic< 10 <alternative< 7 <motion< 4 <paradigm< 1
 stroboscopic  10  alternative  7  motion  4  paradigm  1
|digital| 10 digital 10
di* 5 dig* 5 digi* 5 digit* 5
*gital 10 *ital 10 *tal 10 *al 10
```

Figure 2. FiBRE's structured representation of the annotation presented in Section 2. The representation is composed by an unique identifier (PubMed identifier, gene name, sentence number, place in the sentence, type of annotation), the original type of annotation, and a list of features and their values.

To create the models and classify the annotations, we used Bow, a library that performs statistical text classification using one of several different classification methods [12]. We tested the different classification methods provided by Bow. All of them gave similar results, but the Probabilistic Indexing classification method achieved better performance in both time and accuracy. Thus, the results presented on this paper were obtained using this classification method with forty different 60/40 training/test set splits.

The same annotation can have different classification scores, one for each time the annotation was used in the test set. Thus, to rank the putative misannotations FiBRE has to assign a single confidence score to each annotation based on its multiple classification scores. FiBRE defined the confidence score of each putative misannotation as the minimum of all its classification scores. The reason for selecting the minimum is the same reason why FiBRE only selects the annotations that were never regular, i.e. FiBRE aims at identifying annotations that are consistently irregular independently from the training/test split used. Thus, by selecting the minimum FiBRE is assigning a higher confidence score to the annotations with stable classification scores. Other aggregate functions, besides the minimum, could also be used, such as the maximum, medium or mean.

After assigning a confidence score to each putative misannotation, we can tune the precision and recall of FiBRE by filtering the putative misannotations based on a confidence threshold, i.e., FiBRE returns only the putative misannotations with a confidence score higher than the confidence threshold used. The idea is that the probability of a putative misannotation being a Text Detective error should be proportional to its confidence score. This means that as we increase the confidence threshold we may skip some misannotations but we will improve the precision of FiBRE.

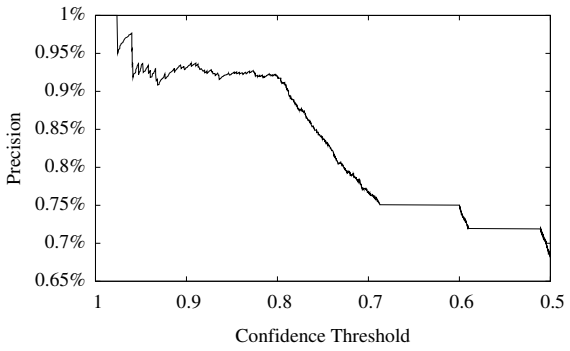

Figure 3. Precision of FiBRE varying the confidence threshold.

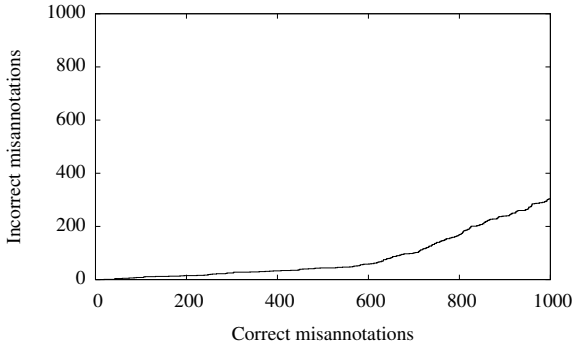

Figure 4. Correct versus incorrect misannotations returned by FiBRE.

4 Results and Discussion

FiBRE classified the annotations recognized by Text Detective from 17,585 abstracts. FiBRE has identified a total of 4,736 putative misannotations from 59,088 gene annotations generated by Text Detective. From these 4,736 putative misannotations, curators manually evaluated the 1,762 gene annotations. Almost all putative misannotations with confidence scores above 0.7 were curated, except 48 of them that were excluded by the curators because they could not take an accurate decision about their correctness.

Table 1 presents some of the annotations that FiBRE predicted to be misannotations. For example, FiBRE identified correctly the misannotation of *QT* because it is surrounded by words that are unusual to be found near a gene, such as *interval*. On the other hand, FiBRE identified incorrectly the misannotation of *GCDH* because it is near the word *deficiency* that is normally near diseases.

Precision measures the fraction of putative misannotations identified by FiBRE that curators confirmed as being real misannotations. Table 2 shows the precision of these annotations. For example, for a confidence threshold of 0.8 only 49 of the 617 putative misannotations evaluated were not curated as being correct misannotations, which corresponds to a precision of 92.1%. Figure 3 represents a plot of the precision obtained by FiBRE over different confidence thresholds. The plot shows that FiBRE has high precision (higher than 90%) for all confidence thresholds above 0.8. Moreover, the precision is almost constant (about 93%) for thresholds between 0.8 and 0.95, which shows the reliability of FiBRE for thresholds above 0.8.

Threshold	Putative	Curated	Correct	Precision	Recall
0.9	223	218	203	93.1%	1.8%
0.8	622	617	568	92.1%	4.8%
0.7	1345	1297	995	76.7%	8.7%
0.6	2429	1414	1060	74.9%	15.4%
0.5	3971	1762	1203	68.3%	22.9%

Table 2. For different confidence score thresholds, this table presents: the number of putative misannotations identified by FiBRE, how many were manually curated, how many were curated as being real misannotations, and the correspondent precision and estimated recall values.

Recall measures the fraction of misannotations generated by Text Detective that FiBRE was able to identify. We can only estimate recall because it was unfeasible to manually evaluate all the gene annotations recognized by Text Detective, as well as the annotations filtered by FiBRE for confidence thresholds below 0.7. To estimate the recall we assume that 20% of the total gene annotations produced by Text Detective are misannotations, as claimed by their authors. Thus, the recall was estimated by the number of correct putative annotations over $59,088 \times 0.2$, and then multiplied by the rate of curation for that confidence threshold, since not all the annotations were curated. Table 2 shows that the filtering performed by FiBRE achieves almost a recall of 5% for a precision of 92.1%. The number of putative misannotations increases exponentially with the threshold, but the number of errors has only a small linear increase, as shown in Figure 4. This also demonstrates that the performance of FiBRE is linearly proportional to the confidence threshold and to the number of putative misannotations identified.

The features that most influenced the classification method to decide the type of the annotation were the following:

```
*tor recep* |receptor| >protein> <expression< protein
*is  canc*  |disease|  >patient> <patient<    cancer
```

The features in the first line can be easily recognized as being important to classify an annotation as gene and the

ones in the second line as non-gene. This shows the ability of FiBRE to automatically identify features that are clearly significant to filter misannotations.

5 Conclusions

Motivated by the vast amount of publications, text-mining systems have been developed to minimize the effort of curators and to help researchers keep up with the progress in a specific area of the biological sciences. However, existing text-mining tools are yet far from reaching performance levels compared to those obtained in other areas. A text-mining tool can only perform well when it is identifying the bioentities correctly, since errors in the recognition are propagated to the text-mining process.

To improve the precision of bioentity recognition tools we developed FiBRE, a system capable of filtering errors made by rule-based named bioentity recognition systems, such as Text Detective. In this article we show that FiBRE was able to identify about 5% of the total Text Detective misannotations, with a precision of 92.1%, and requiring minimal human effort, since it is fully automated and uses only the results of Text Detective. However, FiBRE is only effective when there is a substantial amount of accurate annotations available, otherwise the classification method will be unable to find out the features that characterize the entities. The annotations returned by Text Detective have a precision of about 80%, which was shown to be sufficient to effectively apply FiBRE.

In future work we intend to improve FiBRE by extending the method to all the bioentities returned by Text Detective (and not only genes), thus filtering errors from all kinds of bioentities. We also intend to improve FiBRE by adjusting the parameters of the classifiers for maximum performance; by using efficient voting strategies with different training sets and multiple classifiers; by integrating external domain knowledge and thus generating new features; and by using an individual classifier for each bioentity. We also intend to use probabilistic models, such as a Bayesian network, to decide whether a given annotation is correct or not based on all classification scores instead of using only the minimum score.

6 Acknowledgments

This work was supported by FCT, through the project PTDC/EIA/67722/2006, the Multiannual Funding Programme, and the PhD grants SFRH/BD/42481/2007 and SFRH/BD/36015/2007.

References

[1] C. Blaschke, L. Hirschman, and A. Valencia. Information extraction in Molecular Biology. *Briefings in BioInformatics*, 3(2):154–165, 2002.

[2] H. Chun, Y. Tsuruoka, J. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. *Pac Symp Biocomput*, 2006.

[3] K. Cohen and L. Hunter. Getting started in text mining. *PLoS Comput Biology*, 4(1), 2008.

[4] F. Couto, T. Grego, R. Torres, P. Sánchez, L. Pascual, and C. Blaschke. Filtering bioentity recognition errors in bioliterature using a case-based approach. In *BioLINK SIG, ISMB/ECCB*, 2007.

[5] F. Couto and M. Silva. *Advanced Data Mining Techonologies in Bioinformatics*, chapter Mining the BioLiterature: towards automatic annotation of genes and proteins. Idea Group Inc., 2006.

[6] S. Dickman. Tough mining. *PLoS Biology*, 1(2):144–147, 2003.

[7] W. Hersh, R. Bhuptiraju, L. Ross, P. Johnson, A. Cohen, and D. Kraemer. TREC 2004 genomics track overview. In *Proc. of the 13th Text REtrieval Conference*, 2004.

[8] W. Hersh, C. A. M., R. P., and R. H. K. TREC 2006 genomics track overview. In *Proc. of the 15th Text REtrieval Conference*, 2006.

[9] L. Hirschman, J. Park, J. Tsujii, L. Wong, and C. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561, 2002.

[10] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1, 2005.

[11] M. Krallinger and A. Valencia. Text-mining and information-retrieval services for molecular biology. *Genome Biology*, 6(7):224, 2005.

[12] A. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/∼mccallum/bow/, 1996.

[13] D. Rebholz-Schuhmann, H. Kirsch, and F. Couto. Facts from text - is text mining ready to deliver? *PLoS Biology*, 3(2):e65, 2005.

[14] H. Shatkay and R. Feldman. Mining the biomedical literature in the genomic era: An overview. *Journal of Computational Biology*, 10(6):821–855, 2003.

[15] A. Spencer. *Morphological theory*. Oxford: Blackwell, 1991.

[16] J. Tamames. Text detective: a rule-based system for gene annotation in biomedical texts. *BMC Bioinformatics*, 6(Suppl 1):S10, 2005.

[17] B. Wellner. Weakly supervised learning methods for improving the quality of gene name normalization data. In *Proc. of the workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, 2005.

[18] A. Yeh, L. Hirschman, and A. Morgan. Evaluation of text data mining for database curation: Lessons learned from the KDD challenge cup. *Bioinformatics*, 19(1):i331–i339, 2003.