Filtering Bioentity Recognition Errors in Bioliterature using a Case-based Approach

Francisco M. Couto [a]*, Tiago Grego [a], Rafael Torres [b], Pablo Sanchez [b], Leandro Pascual [b], Christian Blaschke [b]

[a] Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal
[b] Bioalma, Tres Cantos, Spain

ABSTRACT

Motivation: Identification of entities in biomedical literature is a mandatory step for systems attempting accurate information retrieval and extraction. Two main approaches are used by these systems: rule-based and case-based approaches. The rule-based approach relies on rules inferred from patterns identified from the text by an expert. The case-based approach relies on a predefined set of texts previously annotated by an expert.

Results: This paper presents FiBRE, a technique for filtering errors made by a rule-based system, by validating its output using a case-based approach. The results were promising since FiBRE was able to detect 17 misannotations of genes with a precision of 100% in a set of almost 1,000 documents. FiBRE is completely automated and can be applied to any entity recognition system.

Availability: On demand.

Contact: fcouto@di.fc.ul.pt

Keywords: Named bioentity recognition, Rule-based and Case-based approaches.

1 INTRODUCTION

Text Mining generally concerns the process of extracting relevant and non-trivial information and knowledge from unstructured text, usually a collection of documents (Hearst, 1999). One target application of Text Mining is the BioLiterature, from where details of experimental results can be automatically extracted. However, the development of efficient text-mining techniques specific to BioLiterature is a recent research topic. As a result, the observed performance of text-mining tools in BioLiterature has been much lower than in other areas such as news text (Yeh *et al.*, 2003; Hersh *et al.*, 2004; Hirschman *et al.*, 2005).

The main problem in BioLiterature is coping with the lack of a standard nomenclature for describing biologic concepts and entities. In BioLiterature, 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). Genes, whose name is a common English word, are frequent, which makes it difficult to recognise biological entities in the text. The information to be extracted is also more complex. It is almost impossible to derive a rule without having a significant number of exceptions (Dickman, 2003; Rebholz-Schuhmann *et al.*, 2005).

Correctly identifying entities and concepts that are mentioned in a text is a mandatory step for systems attempting accurate information retrieval and, especially, information extraction tasks (Krallinger *et al.*, 2005). This paper presents FiBRE (Filtering Bioentity Recognition Errors), a technique for automatically filtering the errors made by systems that automatically recognize bioentities in biomedical texts. The idea is to automatically learn features that characterise different sets of annotations and then use these features to re-classify the annotations. By annotation, in this paper, we mean the recognition of a bioentity in a piece of text. The learning process can be performed by statistical classification methods that use a training set to create a model that can then be used to classify a test set. FiBRE proposes to use as the training set samples of at least two different sets of annotations, for example gene and non-gene annotations. After creating the model, FiBRE classifies the remaining annotations (test set) to check whether they maintain their original category or not. This technique assumes that annotations which systematically change category are potential errors. FIBRE requires minimal human intervention since the training sets are not manually created.

The remainder of this paper is organised as follows. Section 2 introduces the state-of-the-art approaches used in the named entity recognition field. 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 APPROACHES

Most state-of-the-art text-mining systems use a rule-based or a case-based approach for retrieving information from the text (Leake, 1996; Couto and Silva, 2006).

The rule-based approach relies on rules inferred from patterns identified from the text by an expert. The rules represent, in a structured form, the knowledge acquired by experts when performing the same task. The expert analyses a subpart of the text and identifies common patterns in which the relevant information is expressed. These patterns are then converted to rules to identify the relevant information in the rest of the text. The main bottleneck of this approach is the manual process of creating rules and patterns. Besides being time-consuming this manual process is, in most cases, unable to devise from a subpart of the text the set of rules that encompass all possible cases.

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 (Tamames, 2005). 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

*to whom correspondence should be addressed

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 often 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 case-based approach relies on a predefined set of texts previously annotated by an expert, which is used to learn a model for the rest of the text. Cases contain knowledge in an unprocessed form, and they only describe the output expected by the users for a limited set of examples. The training set is built based on a subpart of the text and on the expected output that should be returned by the text-mining system. The system uses the training set to create a probabilistic model that will be applied to the rest of the text. The main bottleneck of this approach is the selection and creation of a training set large enough to enable the creation of a model accurate for the rest of the text.

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 not 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 results of the rule-based systems to automatically create the training sets, i.e. it is based on weakly supervised learning approaches that were recently tested for identifying gene mentions in text (Wellner, 2005; Chun *et al.*, 2006).

3 FIBRE

This section describes FiBRE in detail and how it was implemented to filter the annotations given by Text Detective.

3.1 Prerequisites

FiBRE can only be applied to rule-based named entity recognition systems that produce annotations at least of two different categories. FiBRE can also be applied to case-based systems, but we think that it would be much less effective than in rule-based systems since in the bottom line we would be applying the same technique twice. The different categories are required to automatically create training sets that contain the features used to differentiate the annotations.

3.2 Input

FiBRE receives two sets of annotations that the named entity recognition system classified in two different categories. In our experiment, the annotations given by Text Detective were split in two categories: one containing the gene annotations, and other containing the remaining non-gene annotations (chemical compounds, drugs, diseases, symptoms).

Each annotation given by Text Detective was composed by the PubMed identifier and the location where the bioentity was recognised within the abstract.

3.3 Output

The output is the list of given annotations that FiBRE classified in a different category. In our experiment, the output is the gene annotations given by Text Detective that FiBRE classified as non-genes and the non-gene annotations given by Text Detective that FiBRE classified as genes.

Each annotation that FiBRE returns is accompanied by a confidence score provided by the used classification method.

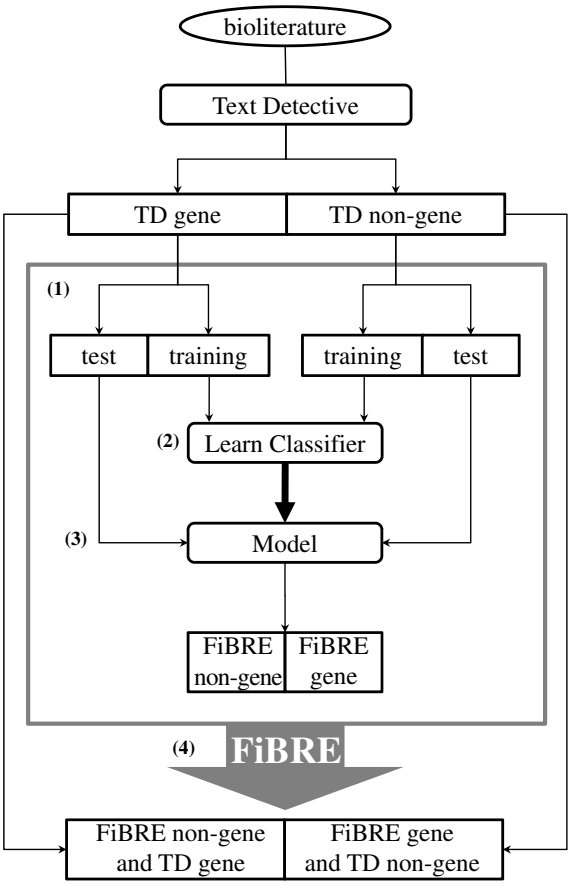

Fig. 1. The named bioentity recognition system (Text Detective) identifies two different categories of annotations (TD gene and TD non-gene) from the BioLiterature. FiBRE splits each of the two sets of annotations in two sets (training and test). The training sets are used for learning the statistical classifier that creates a model, which is used to classify the test sets of both categories. The result is a two sets of annotations, one that FiBRE classified as being gene annotations (FiBRE gene) and the other as being non-gene annotations (FiBRE non-gene). The steps above are executed multiple times with different training/test set splits to include each annotation in the test set at least once. In the end, we have two set of putative misannotations, the ones that are in the TD gene set and in the FiBRE non-gene set, and the ones that are in the TD non-gene set and in the FiBRE gene set.

3.4 Procedure

Figure 1 represents an outline of the main steps of FiBRE that are described below.

In the first step, both categories of annotations are spread in two sets: training and test set.

The second step uses the training sets of both categories to create a model using a statistical classification method.

Table 1. Examples of the annotations returned by FiBRE. The bioentities recognized by Text Detective are inside brackets.

PMID	Sentence	TD Category	FiBRE Category	Score
1693949	Prokaryotic DBH expression yielded a 65-kilodalton DBH-immunoreactive peptide that differed from eukaryotic adrenal DBH only in N-linked, <endoglycosidase F-sensitive glycosylation> in the latter.	gene	nongene	0.96
11211125	The <potassium channel blockers clotrimazole> and tetrapentylammonium (TPeA) inhibited I(Ks) with a lower potency than I(K).	gene	nongene	0.94
1900384	<Plasma high-density-lipoprotein cholesterol> (HDL-C) at the end of each dietary period was not significantly different but the midpoint values were lower by 12.5% on the lower LA diet and 7.3% on the higher LA diet	gene	nongene	0.94
11595829	PJS and CNC share manifestations with Cowden syndrome (or Cowden disease) (<CS>, OMIM#158350) and Bannayan-Riley-Ruvalcaba syndrome (BRR, OMIM#153480)	gene	nongene	0.93
7960690	Exercise should be an integral part of the treatment in non-insulin-dependent (<NIDDM>) diabetic patients, yet most of these patients' performance is low, mainly because of their obesity and concomitant macrovascular disease.	gene	nongene	0.91
1693749	We hypothesise that MeCP2 normally binds methylated DNA in the context of chromatin, contributing to the long-term repression and <nuclease-resistance> of methyl-CpGs.	gene	nongene	0.90

The third step uses the model to classify the test sets of both categories.

The steps above are iterated several times to include each annotation multiple times in the test set by using different training sets. Therefore, in the end we have multiple FiBRE classifications for each annotation.

Step four selects the annotations that were consistently classified by FiBRE in a category different from its original one. This means that only the gene annotations given by Text Detective that were always classified by FiBRE as non-gene are considered to be Text Detective errors.

3.5 Implementation

Each annotation was represented by a set of features, including the words that compose the bioentity recognised by Text Detective, its surrounding words, and their suffixes and prefixes. The list of features used to represent the annotations have different weights according to their distance to the bioentity recognised.

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 (McCallum, 1996). 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.

FiBRE was implemented with Perl scripts that receive the annotations from Text Detective, represent them as a set of features, add them to Bow, perform the classification several times, and select the annotations that were never classified on its original category.

4 RESULTS

We tested FiBRE with the annotations recognised by Text Detective in a set of 969 abstracts. Then, two curators of Bioalma evaluated the gene annotations given by Text Detective that FiBRE classified as non-gene annotations. Table 2 shows the precision of these annotations. For example, from these annotations there were only 17 annotations with a confidence score higher than 80% and all of them were considered to be non-gene annotations by the curators. This means that FiBRE had 100% precision for confidence scores higher than 80%.

Table 2. Precision of FiBRE.

Confidence	Precision	Misannotations
80%	100%	17
75%	91%	74
70%	82%	104
65%	78%	113
60%	72%	125
55%	65%	142
50%	58%	158

FiBRE predicted a total of 289 misannotations from 6,944 gene annotations identified by Text Detective in the 969 documents. The number of misannotations detected by FiBRE goes from 158 (confidence score of 50%) to 17 (confidence score of 80%). This represents 10% and 1% of the total Text Detective misannotations, assuming that 20% of the 6,944 annotations were incorrect as claimed by the authors of Text Detective. Since it was unfeasible to evaluate all the 6,944 gene annotations manually we can only give this estimate of recall.

In Table 1 we present some of the annotations that FiBRE has correctly identified as errors. Given for example the following sentence from the abstract with the PubMed identifier 11211125:

The potassium channel blockers clotrimazole and tetrapentylammonium (TPeA) inhibited I(Ks) with a lower potency than I(K).

Text Detective has annotated *potassium channel blockers clotrimazole* as a gene, and based on a set of features identified on the sentence, FiBRE was able to detect this annotation as an error. The word *clotrimazole*, as well as its sufixes (e.g. *-zole*, *-azole*) and prefixes (e.g. *clo-*, *clotr-*), were important features used to detect the missannotation. The presence of *tetrapentylammonium* near the bioentity was also important to classify it as non-gene.

By a brief evaluation of non-gene annotations given by Text Detective that FiBRE classified as gene annotations we concluded that the precision was much lower than in the previous case (less than 50%). This was expected, since the non-gene annotations used for training correspond to an heterogeneous set of bioentities. Therefore, it was hard for the classification method to identify relevant features that could characterise non-gene annotations.

5 CONCLUSION

Text-mining systems have been used to minimize the effort spent on automatically identifying the facts and the evidence texts in BioLiterature. However, existing text-mining tools do not always provide what the curators want. On the contrary, they spend a large amount of their time finding the right information. A text mining tool can only perform well when it is identifying the bioentities correctly. Errors in the recognition of entities are propagated to the text mining process.

This paper presented FiBRE, a case-based technique capable of filtering errors made by rule-based named bioentity recognition systems, such as Text Detective. Using only the results of Text Detective, FiBRE was able to identify annotation errors, with high precision, and requiring minimal human effort, since it is fully automated. FiBRE can be extended and implemented to all the bioentities (not only genes), thus filtering errors from all kinds of bioentities.

Despite its success, the approach proposed in this paper also has its limitations. It 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 characterise each category. The precision of the annotations returned by Text Detective is about 80%, which was clearly sufficient to effectively learn the classifiers.

In future work we 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 generating new features from dictionaries of biological terms; by integrating external domain knowledge; and by using an individual classifier for each bioentity.

REFERENCES

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

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

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

Hearst, M. (1999). Untangling text data mining. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*.

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

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

Krallinger, M., Erhardt, R. A.-A., and Valencia, A. (2005). Text-mining approaches in molecular biology and biomedicine. *Drug Discov Today*, **10**(6), 439–445.

Leake, D. (1996). *Case-Based Reasoning: Experiences, Lessons, and Future Directions*. AAAI Press/MIT Press.

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

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

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

Wellner, B. (2005). 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*.

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