BIOINFORMATICS ORIGINAL PAPER
Vol. 24 no. 18 2008, pages 2086–2093
doi:10.1093/bioinformatics/btn381

Data and text mining

Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users

Hagit Shatkay[1,*], Fengxia Pan[1], Andrey Rzhetsky[2,3] and W. John Wilbur[4]

[1]The Computational Biology and Machine Learning Lab, School of Computing, Queen's University, Kingston, Ontario, Canada, [2]Department of Medicine, [3]Department of Human Genetics, Computation Institute, and Institute for Genomics and Systems Biology, University of Chicago, Chicago, IL and [4]National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD, USA

Received on May 25, 2008; revised on July 17, 2008; accepted on July 19, 2008

Advance Access publication August 20, 2008

Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: Much current research in biomedical text mining is concerned with serving biologists by extracting certain information from scientific text. We note that there is no 'average biologist' client; different users have distinct needs. For instance, as noted in past evaluation efforts (BioCreative, TREC, KDD) database curators are often interested in sentences showing *experimental evidence* and *methods*. Conversely, lab scientists searching for known information about a protein may seek facts, typically stated with high confidence. Text-mining systems can target specific end-users and become more effective, if the system can first identify text regions rich in the type of scientific content that is of interest to the user, retrieve documents that have many such regions, and focus on fact extraction from these regions. Here, we study the ability to characterize and classify such text automatically. We have recently introduced a multi-dimensional categorization and annotation scheme, developed to be applicable to a wide variety of biomedical documents and scientific statements, while intended to support specific biomedical retrieval and extraction tasks.

Results: The annotation scheme was applied to a large corpus in a controlled effort by eight independent annotators, where three individual annotators independently tagged each sentence. We then trained and tested machine learning classifiers to automatically categorize sentence fragments based on the annotation. We discuss here the issues involved in this task, and present an overview of the results. The latter strongly suggest that automatic annotation along most of the dimensions is highly feasible, and that this new framework for scientific sentence categorization is applicable in practice.

Contact: shatkay@cs.queensu.ca

1 INTRODUCTION

Biomedical text mining is the focus of much current research (Cohen and Hunter, 2005; Krallinger and Valencia, 2005; Shatkay, 2005), including work on information extraction from the biomedical literature (Blaschke *et al.*, 1999; Craven and Kumlien, 1999; Friedman *et al.*, 2001; Tanabe and Wilbur, 2002), as well as on information retrieval and text categorization (Hersh *et al.*, 2006; Raychaudhuri *et al.*, 2003; Shatkay *et al.*, 2000). Information extraction efforts concentrate primarily on identifying bio-entities (mostly genes and proteins) and relationships among them, while current efforts on information retrieval, with a few exceptions, aim at identifying documents for specific database curation tasks and categorization of papers into various ontological entries (Hersh *et al.*, 2006; Yeh *et al.*, 2003). However, the fact that a gene is mentioned, and even information about it is provided, does not necessarily imply that the information is reliable or useful in satisfying the scientist's information need (Krauthammer *et al.*, 2002; Light *et al.*, 2004; Medlock and Briscoe, 2007).

Moreover, the aforementioned scientist is not a single entity with one simple type of information need. Biomedical database curators (who integrate information from the literature into current public databases such as SwissProt or FlyBase) face a very different task from a researcher looking for information about a certain gene, or a physician looking for the latest drug developments pertaining to a certain disease.

Our broad idea, which is the basis for this work, is that the contents of scientific statements can be characterized along certain general dimensions. In turn, the characteristics of each phrase, sentence or paragraph along these dimensions can help to determine whether the text is useful to a particular user with specific information needs. For instance, a scientist looking for all the information published about a certain gene, may be satisfied by obtaining all the papers or all the sentences mentioning this gene. In contrast, a database curator for the FlyBase or the Mouse Genome Informatics databases, who is looking for experimental evidence that the gene was expressed under certain conditions, would only be satisfied with sentences discussing *experimental evidence* and stating with *high confidence* that the gene was indeed expressed under the reported conditions. As was discussed in the BioCreative II workshop (2007), and pointed out in other previous evaluation efforts [TREC Genomics (Hersh *et al.*, 2006) and the KDD cup (Yeh *et al.*, 2003)], there is much interest in identification and extraction of *methods* and *evidence* passages from the literature, in support of current scientific research and database curation. Thus, by identifying candidate documents, and regions within them, that are rich in experimental evidence and methodological details, and by then focusing extraction efforts on these regions, a text mining system can provide curators with candidate sentences that both

*To whom correspondence should be addressed.

describe the desired phenomenon (e.g. protein–protein interaction or the expression of a gene), as well as *bear the evidence* for the phenomenon or describe the methods by which the phenomenon was identified.

We note that the general idea of reducing the document search space for a specific domain in order to improve retrieval and extraction is not new. Domain-specific document collections and search engines, such as Google Scholar and PubMed, both already effectively demonstrate that a search in a target-focused database has clear advantages with respect to search within a broader and more general space. As such, we note that our goal here is not to reprove this idea, but rather to take a step toward enabling the creation of well-focused subsets of biomedical text that have certain properties. Namely, our goal is to demonstrate the ability to categorize text along certain critical dimensions of interest.

In an earlier paper (Wilbur *et al*., 2006), we introduced criteria for characterizing statements made in the literature along several dimensions, namely: *focus* (e.g. methodology, scientific discovery or generic), *polarity* (positive versus negative statement), level of *certainty*, type of *evidence* and *trend* (increase or decrease in certain measurement). Classifying text along these dimensions provides tags to each text fragment. The utility of the text as a source for certain types of knowledge can be evaluated based on its tags, and specific curation-related or discovery-oriented queries can be supported accordingly. For instance, fragments with methodology statements, or those bearing affirmative scientific statements supported by experimental evidence are of high utility for database curators (Yeh *et al*., 2003). Earlier work on annotation of scientific text (Langer *et al.*, 2004; Mizuta and Collier, 2004; Teufel *et al*., 1999), which we have surveyed in detail before (Wilbur *et al*., 2006), focused on the partition of text into zones, based on types of discourse and components of scientific argumentation (e.g. *background, aim*). In contrast, our approach introduces five general dimensions, along which each sentence or sentence fragment within the text can be characterized, regardless of its role or its zone within the article.

Based on the new criteria, a major annotation effort took place, in which 10 000 sentences taken from full-text articles and from biomedical abstracts were annotated by a team of eight *independent*, well-trained annotators (none of them is an author of this article). Each sentence was annotated by three different members of the annotation team. The multiplicity of annotations per sentence serves to ensure the quality and the consistency of the annotation. As explained in Section 2, sentences were broken into individual statements (which we call *fragments*) and each statement received its own annotation tags. The resulting large set of annotated statements (sentence fragments) is then used to train and test machine learning methods to classify scientific statements along the multiple dimensions we have defined.

Here, we report our first experiments and results of developing, training and testing machine learning classifiers using the corpus. We present results from this extensive experiment, demonstrating that classification of scientific text along the specified dimensions is indeed feasible. As such, our categorization scheme can indeed be used as a basis for future retrieval and extraction engines.

We note that the categorization task is not trivial, as it has two special characteristics: (1) it is multi-dimensional, that is, it simultaneously categorizes fragments along multiple dimensions; and (2) the categories assigned along each of the dimensions may not be independent of each other. While these two aspects are further discussed in the following sections, we observe here that multi-dimensional categorization, while sharing some aspects with the problem of multi-label categorization (Boutell *et al*., 2004; Ghamrawi and McCallum, 2005) is not the same problem under a different name. Multi-label categorization typically involves a single set of class labels (along a single dimension such as diseases or news categories), and is concerned with assigning the same item into more than one class. Multi-dimensional categorization is concerned with categorizing items using multiple *sets of classes*, where for each dimension a different set of possible classes is defined. The categorization along each of the dimensions can, in turn, be multi- or single-labeled in and of itself.

The rest of the article discusses our categorization scheme, the text-representation methods used to accommodate it, the classifiers and the results obtained by using them. Our findings show that automatic classification of scientific text along the proposed dimensions can be performed effectively. Future directions and applications are outlined.

2 DATA: THE CORPUS AND ITS REPRESENTATION

The whole corpus (a subset of which is used in the experiments as described below) consists of 10 000 sentences[1] selected at random from both full-text articles from a wide variety of biomedical journals, and from biomedical abstracts—sampled from PubMed.[2]

Each of these sentences was annotated by three independent annotators, all of whom have an advanced degree in the biomedical sciences and a good mastery of the English language, who as such represent well the high-end of curators and annotators typically working in biomedical curation. A total of eight annotators divided into subgroups of three worked on the corpus. We next describe the corpus and the datasets used in our experiments.

2.1 The annotated corpus

A full discussion of the annotation scheme and inter-annotator agreement under this scheme was provided in an earlier publication (Wilbur *et al*., 2006). To make this current report self-contained, we briefly review the scheme here, but the reader is referred to the earlier paper for details. Each statement in the corpus (where a statement may be a sentence or just a fragment of a sentence, as described below) was characterized and marked-up along the following dimensions:

- *Focus:* the type of the information conveyed by the statement. Focus can be: *Scientific* (*S*)—discussing findings and discovery; *Generic* (*G*)—stating general knowledge, clarifying the structure of the paper, *etc.*; *Methodology* (*M*)—describing a procedure or a method.
- *Polarity:* indicates whether the statement is made in the affirmative (*P*), or in the negative (*N*).
- *Certainty:* the degree of certainty regarding the validity of the statement, on a scale of 0–3.

[1]The sentence corpus will be provided upon request.
[2]http://www.ncbi.nlm.nih.gov/entrez/

The binding of both forms of β-catenin to CBP is completely inhibited by ICG-001 (Fig. 3B Top, lane 4). ***1SP3E3-***

We demonstrate that ICG-001 binds specifically to CBP ***1SP2E3*** but not the related transcriptional coactivator p300, ***2SN2E3***

Fig. 1. Two annotated sentences. The first constitutes a single fragment. The second is fragmented where the statement's polarity changes from positive (*P*) to negative (*N*).

- *Evidence:* indicates the type of evidence supporting the statement: *E0*—no stated evidence or a stated *lack of evidence*; *E*1—mentions of evidence with no explicit reference; *E*2—the statement is backed by a reference to a supporting publication; *E*3—experimental evidence is directly given in the text.
- *Direction/Trend:* indicates whether an *increase* (+) or a *decrease* (–) in a specific phenomenon, finding or activity is reported in the statement. This field is empty if the text does not provide any indication of a trend.

The scope of the text units to which such tags are assigned, are *fragments* within a sentence. The annotator breaks the sentence when there is a change along any of the five dimensions listed above (e.g. a statement's polarity changes from positive to negative). Each annotation tag starts with the ordinal number of the fragment within the sentence. Figure 1 shows two examples tagged with annotations, where the first did not require fragmentation and the second did.

2.2 The datasets

While the complete corpus consists of 10 000 sentences, as there is a varying degree of inter-annotator agreement along the five annotation dimensions (Wilbur *et al.*, 2006), our current study focuses only on sentences for which *all* three annotators agreed on their annotation tags.[3] We do not require agreement along all the dimensions simultaneously, but rather examine each dimension separately. For instance, to train and test a classifier that distinguishes among the different types of Focus, we use the dataset *Frag_F*, whose fragments' Focus was agreed upon by all three annotators. The classifier, in turn, is tested only on fragments with this type of inter-annotator agreement.

The requirement for inter-annotator consensus provides a unique and reliable tag assignment to each fragment, and supports a clean training/test set, but these advantages come at a price: inter-annotator consensus on a fragment may imply that the fragment is easier to automatically classify, resulting in an easier classification task for a machine learning method. However, along several dimensions and classes only a small number of examples have inter-annotator consensus, making for scarce training and test data and for a *more difficult* machine learning task. As these two effects tend to balance each other, and because even under the consensus requirement our dataset includes thousands of examples, we believe that for this first study of a new approach, the advantages of using consensus as a baseline justifies the possible shortcomings. To investigate the consequences of this decision, we performed experiments with noisier data (using tags based on *majority* rather than consensus) with similar results (not shown here for lack of space; See Pan, 2006). We are currently studying ways to make optimal use of the whole dataset.

[3]We note that there is no absolute notion of truth in tagging text by Evidence, Certainty, Focus, etc. It is the reader's decision, made in practice by database curators, and in this study by our annotators. As the probability of chance agreement among three annotators, even on a binary-valued annotation, such as polarity, is very low (≤0.125), we view consensus among annotators as an indication that this annotation is very likely to be acceptable to biologists; and as such use it as our measure of 'truth' in this study.

Table 1. The number of sentences and of fragments, for which there is complete agreement in annotation, along each dimension

	Foc. & Ev.	Focus	Evidence	Certainty	Polarity	Trend
Dataset name	*Frag_FE*	*Frag_F*	*Frag_E*	*Frag_C*	*Frag_P*	*Frag_T*
Sentences	1977	4068	2964	5644	6430	5907
Fragments	2109	4447	3133	5992	6945	6330
No. of terms selected	F: 600 + E: 200	1500	1500	100	600	100

The 'Foc. & Ev.' column counts agreement along both Focus and Evidence dimensions used for training a classifier that exploits their dependency. The bottom row lists the optimal number of terms selected to represent text along each of the respective dimensions.

Table 1 shows the number of sentences and respective fragments on which all three annotators agreed, for each dimension. As discussed later, annotations of Focus and Evidence are not independent of each other. We thus train a Maximum Entropy classifier that exploits the correlation between the two dimensions. This classifier is trained on fragments for which all three annotators agreed on both the Focus and the Evidence tagging, denoted as *Frag_FE.*

2.3 Preprocessing: term definition and feature selection

The first step we take in text categorization is the representation of the text as a weighted term vector. This representation requires first, to define and extract *terms* from the text (which may yield a very large set of terms), and second, to reduce the term space and select a representative subset, T, of terms that optimize classification performance. The two steps are described subsequently.

2.3.1 Term definition We consider three types of terms: individual *words*, *statistical phrases*—which are sequences of up to three consecutive words (bigrams and trigrams), and *syntactical phrases*—grammatically coherent units within the sentence, such as noun phrases and verb phrases. To identify terms we use two main tools: the *MedPost* tokenizer and part-of-speech tagger (Smith *et al.*, 2004), which produces individual words along with their part-of-speech tags, and *YamCha* (Kudo and Matsumoto, 2000)—a text chunker that identifies *syntactical* phrases. Single words, as well as bigrams and trigrams are also extracted and considered as *statistical* phrases.

For categorizing along the Focus, Evidence and Certainty dimensions, text fragments are represented using simple terms consisting of single words, as well as bigrams and trigrams (statistical phrases). In contrast, for the Polarity and the Trend dimensions, our early analysis (on a few hundred fragments) suggested that syntactical structure bears much information. For instance, the word '*not*' in the fragment *'Epidemiological data do not support the link',* directly implies a negative polarity.

However, in the fragment: '*This overproduction is responsible not only for bone formation but also for…*' the term '*not only*' actually supports an affirmative statement. Thus, for the Polarity and Trend dimensions we use single words and *syntactical phrases* (rather than statistical ones).

2.3.2 Term space reduction: Typically, a series of standard text-processing steps includes: Removal of rare terms, e.g. those occurring only once in the whole corpus; *Stemming*—removal of standard suffixes; Stop word removal (removal of uninformative words, such as *articles*, *pronouns* and *prepositions*); and term selection based on the categorization task, e.g. by using the χ^2-test (Yang and Pederson, 1997).

However, our choice of steps varied with the target representation for each specific dimension. (We note that our choice of steps, terms, rules and strategies described here was entirely based on the analysis of a few hundred fragments—about 200 sentences—*predating* the current corpus. As such none of the choices made here is based on the test data.)

In particular, stemming, using the Porter stemmer (Porter, 1980), was applied to terms when representing text for classification along the *Polarity* and *Trend* dimensions, but not along the other three dimensions. An example motivating this choice is the importance of the past tense used when reporting one's own experiments, as in 'We *demonstrated* that…'. The form '*…ed*' indicates that an experiment was performed by the authors; suffix removal is thus likely to impede the correct categorization along the *Evidence* dimension. Similar considerations apply to the *Focus* and the *Certainty* dimensions.

Standard stop-word removal is another common practice. However, we find that some standard stop words are useful indicators for specific types of classes. For instance pronouns such as *we* or *their*, are viewed as stop words with respect to the Trend dimension, but not for the Evidence dimension; in the context of Evidence, these words can distinguish between experimental work done by the authors (evidence type *E3*) and work done by others (evidence type *E1*). Table 2 shows examples of standard stop words and indicates which of those are viewed as stop words along each of the dimensions. A complete list is provided separately (Pan, 2006).

Aside from standard stop words, other terms are discarded from the text for classification along some of the dimensions, based on their part of speech. For instance, pronouns typically do not reflect Polarity or Trend, but are informative with respect to Evidence and Focus. For the complete set of rules of inclusion/exclusion of terms based on their syntactical roles see Pan (2006).

Table 2. Examples of standard stop-words that serve as stop-words along some dimensions (denoted:✓) but are regarded as content-bearing along the other dimensions (denoted: 0). The dimensions are listed as F (focus), P (polarity), C (certainty), E (evidence) and T (trend)

Word	F	P	C	E	T	Word	F	P	C	E	T	Word	F	P	C	E	T
A	✓	✓	✓	✓	✓	may	✓	0	0	0	✓	hence	✓	✓	0	0	✓
always	✓	0	0	0	✓	not	✓	0	0	0	✓	rather	✓	0	✓	✓	✓
perhaps	✓	✓	0	0	✓	their	✓	✓	✓	0	✓	using	0	✓	✓	0	✓

We stress again that the rules and term lists above were all created based on the analysis of a few hundred fragments (about 200 sentences) *predating* the corpus. As such these rules do not over-fit the training and test data used by the classifiers.

After removal of *stop words*, rare terms and term removal based on syntactic forms as described earlier, the number of remaining terms per dimension ranges between 4000 and 10 000. To further reduce this large space, we experimented with several term selection functions (studied by Yang *et al.*). Here, we apply the χ^2 criterion (Yang and Pederson, 1997) and select the highest ranking terms. The number of terms per dimension is shown in the bottom row of Table 1. Once terms are selected, we use simple binary weight vectors to represent the text.

3 CLASSIFICATION METHODS

The multi-dimensional categorization task has unique characteristics. Typically in machine learning, a dataset is classified along a single axis or property (e.g. credit card transactions are partitioned into *fraudulent* versus *legitimate* transactions). In contrast, multi-dimensional annotation simultaneously categorizes each fragment along multiple dimensions. For instance, each fragment is assigned a category along the *Focus* dimension, indicating whether the fragment discusses a scientific topic, (*S*), a method (*M*) or makes a generic statement (*G*). The same fragment is also categorized based on its *Polarity* (whether it makes a positive or a negative statement), and several other dimensions as listed before.

The task not only requires classification along multiple dimensions, but also introduces additional complexity of two kinds: First, along some dimensions a fragment can be assigned more than one class; that is, the task involves multi-label classification (Boutell *et al*., 2004; Ghamrawi and McCallum, 2005). For instance, a statement may be supported by several types of evidence. Specifically, consider the fragment:

> *…the overexpression of phospho-H2Av did not induce G2/M arrest or affect DSB-dependent G2/M arrest (fig. S10) (14,21), **1SN3E23+*

It refers to an experimental figure '*fig. S10*' as well as cites other papers '*(14,21)*.' Thus, the evidence type of the fragment is E23, that is, both *Explicit citation* (denoted as *E2*) and *Explicit evidence* (denoted as *E3*).

Similarly, the sentence:

> *Future structural and functional studies will be necessary to understand precisely how She2p binds ASH1 mRNA and how interactions with She3p influence the formation of a functional localization complex. **1SGP0E0*

poses scientific questions (*how She2p binds ASH1 mRNA* and *how interactions with She3p influence the formation of…*) while also talking about the general state of knowledge, '*Future structural and functional studies will be necessary*'. Therefore, the Focus of the sentence is *SG*, that is, both *Scientific* and *Generic*.

The second kind of introduced complexity is that classification along the different dimensions may not be independent. In particular, most (over 90%) of the fragments that discuss methodology (Focus *M*), describe scientific experiments performed by the authors, with a consequent Evidence type *E3*. Similarly, most (over 55%) of the fragments that make generic statements (Focus *G*) do not support the statements with evidence, implying Evidence type *E0*. It is, therefore, likely that taking advantage of the correlation between these two dimensions in the categorization process, can improve classification results along both the Focus and the Evidence dimensions. This idea is tested in some of our experiments.

To address the special aspects of this classification challenge, specifically along the Focus and the Evidence dimensions, a new classification model based on Maximum Entropy was designed, and experimented with, in addition to standard classifiers. The classification along the other three dimensions, Polarity, Trend and Certainty, is treated as three individual

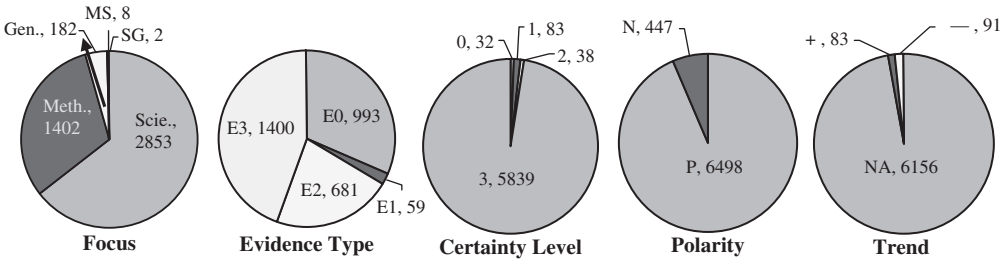

Fig. 2. Distribution of annotation values along each of the dimensions in the dataset. The number of fragments in each class is shown next to the class tag. The dataset shown covers only annotations on which all three annotators were in agreement. Notably, along the certainty, polarity and trend dimensions, almost all fragments are annotated with the highest certainty (3), positive polarity and no trend (neither increase nor decrease in measurement), respectively. The focus and evidence distributions are also skewed, with over a half of the fragments (2853) discussing science, and over a third (1400) providing experimental evidence (*E3*).

text classification tasks and performed separately for each class, with one classifier per dimension[4].

Since the basic classification unit is a sentence fragment with a few words, leading to sparse representation, we use support vector machines (SVMs), which work well on sparse data and typically outperform other text classification methods. We employ the LibSVM implementation (Chang and Lin, 2001), with the radial basis functions kernel and the one-against-one approach for multi-class categorization. Experiments using naïve Bayes classifiers produced similar results (data not shown). We next provide a short overview of the Maximum Entropy model, as it applies to the Focus and the Evidence dimensions.

The Maximum Entropy method (Nigam *et al*., 1999) for model learning is based on the principle that in the absence of prior knowledge, the least informative distribution, i.e. the distribution with the maximum entropy, is preferred. Maximum Entropy classifiers, (similar to naïve Bayes classifiers), estimate the conditional probability of the class label given the text fragment, that is, $p(c|d)$, where d is an input text, and c denotes a class. In contrast to other classifiers, the training data is explicitly used to set certain constraints on the conditional distribution $p(c|d)$. More formally, if $|D|$ denotes the number of training examples, d denotes a text fragment, C denotes the set of all possible classes, c denotes one class ($c \in C$), $C(d)$ is the true class of text fragment d and f_i denotes a feature function which produces one of the features representing a fragment (a *term-weight* is an example of such a feature), the conditional distribution must satisfy the constraint:

$$\frac{1}{|D|}\sum_{d\in D} f_i(d, C(d)) = \sum_{d\in D} p(d)\sum_{c\in C} p(c|d) f_i(c, d).$$

The left side of the above equation denotes the expected value of the feature f_i based on the training data, while the right side denotes the expected value of the feature f_i based on the classification model distribution $p(c|d)$.

The task of learning a classifier from a set of categorized training data, D, is to find, among all the distributions $p(c|d)$ that satisfy the above constraints, the optimal distribution $p^*(c|d)$ that maximizes the conditional entropy, defined as:

$$H(p) = -\sum_{d\in D}\sum_{c\in C} p(d)p(c|d)\log[p(c|d)].$$

The main advantages of the method are: (1) Parameter estimation is computationally simple, and (2) Unlike naïve Bayes, it does not assume conditional independence among features. Moreover, and most important to our current application, the method provides a simple way to introduce constraints and dependencies based on prior knowledge. We utilize the latter to directly incorporate constraints, such as the correlation between the Focus and the Evidence dimensions—discussed earlier—into the learning process.

Explicitly, we train a classifier that assigns a pair of *Focus–Evidence* tag to each fragment. As Focus has 3 possible values (*S*, *M*, *G*) and Evidence has 4 (*E0*, *E1*, *E2*, *E3*), there are 12 possible tag pairs. We define a binary feature value $f_{ij}(d)$ ($1 \le i \le 3$; $1 \le j \le 4$), for each text fragment d, such that $f_{ij}(d) = 1$ if d is tagged by the i-th Focus value and the j-th Evidence value, and $f_{ij}(d) = 0$ otherwise. The constraint imposed is that $E_{\text{training}}(f_{ij}) = E_p(f_{ij})$. That is, the expected co-occurrence statistics of the Focus and Evidence tags produced by the categorization model should be the same as the co-occurrence statistics manifested in the training data.

4 EXPERIMENTS AND RESULTS

4.1 Experimental setting

To test our classification methods we apply a stratified 5-fold cross-validation scheme: the dataset is split into five equal-sized subsets, each with the same class distribution as the whole dataset. Five learning epochs are executed, of which each uses 80% of the data (4-folds) for training and 20% (1-fold) for testing. This evaluation scheme is widely accepted in machine learning (Hastie *et al*., 2003).

We note that, as shown in Figure 2, the distribution of the annotated data is quite skewed. For instance, along the Focus dimension, about 2/3 of the fragments discuss Science, about 1/3 discuss Methodology, a small fraction have a Generic topic and a scant few are tagged with a combination (MS, SG).

While the Evidence categorization is more balanced among evidence types *E0* (no evidence), *E2* (evidence by citation) and *E3* (the majority, direct experimental evidence), only about 2% have evidence of type *E1* (an indirect suggestion of evidence). Much more skewed are the distributions of samples along the other dimensions—Certainty, Polarity and Trend.

The skewed distributions imply that categorization into the underrepresented classes may be inaccurate, as there is not enough data to train on. However, we note that the well-represented classes within this data include categories of much interest for database curators (Hersh *et al.*, 2006; Yeh *et al*., 2003). Namely, *methodology* and *scientific* statements within the Focus dimension, and *experimental evidence* (*E3*), evidence by citation (*E2*) and lack-of-evidence (*E0*) along the evidence dimension. Classifiers that can identify these categories with high precision and *recall* are likely to be of much use. As the results for both Evidence and Focus indicate (Tables 3 and 4), such classifiers were effectively learned from this data.

[4]We thank the anonymous referee for pointing out that Maximum Entropy classifiers that handle dependence in the multi-label case were developed by Ghamrawi and McCallum (2005). Our classifiers were developed independently, for addressing dependence among labels in the multi-dimensional case, at about the same time as described by (Pan, 2006).

Table 3. The results of SVM classification in a 5-fold cross-validation experiment

Category	No. of Fragments	Precision	Recall	*F*-measure	Accuracy
Evidence (FRAG_E)					
E0	993	0.85	0.93	0.89	
E1	59	0.77	0.46	0.57	
E2	681	0.94	0.94	0.94	
E3	1400	0.93	0.89	0.91	
Average		**0.87**	**0.80**	**0.84**	
Weighted average		**0.90**	**0.90**	**0.90**	**0.90**
Focus (FRAG_F)					
Science	2858	0.91	0.98	0.94	
Methodology	1406	0.95	0.86	0.91	
Generic	183	0.94	0.41	0.57	
Average		0.93	0.75	0.83	
Weighted average		0.92	0.92	0.92	0.92
Certainty (FRAG_C)					
0	32	0.71	0.53	0.61	
1	83	0.95	0.63	0.75	
2	38	0.68	0.34	0.46	
3	5832	0.99	1.00	0.99	
Average		**0.83**	**0.62**	**0.71**	
Weighted average		**0.99**	**0.99**	**0.98**	**0.99**
Polarity (FRAG_P)					
P	6498	1.0	1.0	1.0	
N	447	0.96	0.93	0.95	
Average		**0.98**	**0.99**	**0.97**	
Weighted average		**1.0**	**0.99**	**1.0**	**0.99**
Trend (FRAG_T)					
No Trend	6156	0.98	0.99	0.99	
+	83	0.64	0.39	0.48	
−	91	0.66	0.27	0.39	
Average		**0.76**	**0.55**	**0.64**	
Weighted average		**0.97**	**0.98**	**0.97**	**0.98**

Within each of the five dimensions, we show for each of the categories the *Precision*, the *Recall*, the *F*-measure, along with the global measures: average, weighted average and overall *accuracy* of the classification. The number of fragments in each category within the dataset is also listed to emphasize the large variance in the number of samples among the different categories.

It is well known that when the data is skewed, a naïve classification scheme, which assigns every instance to the majority class may be just as accurate as any machine learning method (especially with respect to Certainty, Polarity and Trend). To address this concern we measure not only the overall *accuracy* of each classifier but also its *precision* and the *recall* with respect to each individual dimension and category within dimension, as well as the commonly used F-measure (Van Rijcsbergen, 1979).

Formally, these measures are defined as follows: let $|D|$ be the total number of fragments, and for each category c, *we denote* the number of true positives (fragments correctly assigned to c) as TP_c; the number of false positives (fragment erroneously assigned to category c) as FP_c and the number of *false negatives* (fragments whose true category is c, but were not assigned to it by the classifier) is FN_c. The evaluation measures are then defined as:

$$Precision_c = \frac{TPc}{Tp_c + Fp_c} \quad Recall_c = \frac{TP_c}{TP_c + FN_c} \quad Accuracy = \frac{\sum_{c \in C} TP_c}{|D|}$$

Table 4. The performance of the Maximum Entropy classifiers along the joint Focus and the Evidence dimension on the dataset Frag_FE

Category	No. of Fragments	Precision		Recall		*F*-measure		Accuracy	
		ME_1	ME_2	ME_1	ME_2	ME_1	ME_2	ME_1	ME_2
FOCUS									
S	1420	0.91	0.94	0.96	0.97	0.94	0.96		
M	620	0.90	0.92	0.84	0.90	0.87	0.91		
G	69	0.77	0.83	0.31	0.38	0.44	0.52		
Average		**0.86**	**0.90**	**0.71**	**0.75**	**0.78**	**0.82**		
Weighted average		0.90	**0.93**	0.90	**0.93**	0.90	**0.93**	0.90	**0.93**
EVIDENCE									
E0	696	0.82	0.80	0.91	0.91	0.86	0.85		
E1	53	0.65	0.70	0.28	0.40	0.39	0.51		
E2	493	0.90	0.93	0.87	0.84	0.89	0.88		
E3	867	0.91	0.92	0.87	0.89	0.89	0.91		
Average		0.82	**0.84**	0.74	**0.76**	0.78	**0.80**		
Weighted average		0.87	**0.88**	0.87	0.87	0.87	0.87	0.87	0.87

The dataset contains 1977 sentences (2109 fragments). ME_1 is a basic maximum entropy classifier, while ME_2 incorporates constraints reflecting the correlation between the annotations along the two dimensions. The higher values on the global results (averages and accuracies) are shown in boldface.

where the *F*-measure is calculated as:

$$F_c = \frac{2 \cdot Precision_c \cdot Recall_c}{(Precision_c + Recall_c)}.$$

We also report for each dimension the average *precision*, *recall* and *F*-measure over all the classes, using both a simple average—where all classes have an equal weight (macroaverage), and a weighted average—where the *precision*, *recall* and *F*-measure per class c is weighted proportionally to the respective number of fragments that were annotated as belonging to c (microaverage).

The weighted average provides a measure of the expected performance-per-fragment, in view of the fact that text fragments are not uniformly distributed among the classes.

We report the results of classifying data into the categories along each dimension separately using SVMs. These experiments were performed on each of the datasets *Frag_F, Frag_E, Frag_C, Frag_P* and *Frag_T*. We also show results of classifying the smaller dataset on which all annotators agreed on both the Focus and the Evidence (*Frag_FE*), using the Maximum Entropy classifiers. To this end, we compare the results from using a Maximum Entropy classifier that does not impose dependency constraints between the two dimensions to those obtained when the constraints are imposed.

As stated before, we have shown in early experiments similar results on noisier data where agreement by two or more annotators (rather than by all three) was used (Pan, 2006). These are not reproduced here for the sake of brevity.

4.2 Results

Table 3 shows the results for the classification of the datasets along each of the five dimensions. The *Average* field shows a simple average of each measure along all the categories within each dimension, while the weighted average weighs the measure obtained for each category by the number of fragments within the category, and divides by the total number of fragments. The results, even along the dimensions that show the most skewed annotation distribution,

demonstrate that classification actually improves performance with respect to the naïve approach that would have assigned all instances to the majority category. This is true for all evaluation measures. In particular, the relatively low average F-measure for Trend (0.64) would have been 0.33 under the naïve approach. (That is, if we were to simply assign all fragments to the majority class, namely to the class *No Trend*, the F-measure would be approximately 1 for the *No Trend* class and 0 for both the '+' and the '–' classes, resulting in an average F-measure of 0.33.) Along the Polarity dimension the advantage is even clearer: the naïve approach would produce *precision* and F-measure of 0 for the negative class (as all fragments would simply be assigned to the positive majority class), while the classifier learned from the data, as shown in the table, has high *precision*, *recall* and F-measure for both the positive and negative classes.

Table 4 shows the results obtained when applying Maximum Entropy classifiers to the dataset *Frag_FE*—the fragments on which all three annotators agreed on both Focus and Evidence. These fragments were used for training and testing (in a 5-fold cross-validation setting) a simple Maximum Entropy classifier, (denoted ME_1), as well as one that takes into account the correlation between Focus and Evidence (denoted ME_2). While the dataset *Frag_FE* is much smaller than any of the five datasets used to produce Table 3 (as such the results are not directly comparable to those obtained using SVM), we note that the results obtained from the classifier that uses the correlation between Focus and Evidence (ME_2) improve upon those that do not take the correlation into account (ME_1). This suggests that improvement in classification can be gained by using the interdependencies between the two dimensions.

5 DISCUSSION AND FUTURE WORK

We presented results from a first set of experiments in automatically training and testing machine learning classifiers on a large corpus of sentence fragments, based on a new annotation scheme.

This scheme allows the categorization of text along multiple important dimensions, supporting the identification of information that has been recently pointed out by biologists and biological database curators as highly desirable, including: experimental evidence, indirect evidence, description of methods, positive versus negative statements and others. The manual annotation was performed by multiple independent annotators. Overall high values of the results imply that reliable classification along the suggested dimensions is well within reach.

Clearly, in several areas performance is lower, most notably along the Trend dimension. By manually checking the annotations we noticed inconsistency in the way the annotators applied the annotation guidelines along this dimension, leading to fewer than 100 examples for each of the negative and the positive trends, and performance, as expected, deteriorates.

The dearth of data along certain dimensions can be addressed by utilizing data on which annotators disagreed. While we already conducted experiments using majority annotation as opposed to consensus, we are more interested in a different route, in which we formally model the reliability of each of the annotators, and using the most trustworthy annotations as the basis for training and testing of classifiers. We are currently pursuing this route and expect that using a larger portion of the annotated corpus for training and testing will improve the classification performance.

Another extension to be taken in the near future is the automatic processing of complete documents, by automatically breaking sentences into fragments, and using our classifiers to annotate each fragment.

We believe that the categorization of text along the multiple suggested dimensions, which characterize important aspects of scientific contents, can lead to more accurate extraction and retrieval of information from a large volume of publications. Specifically, being able to retrieve documents that are rich in experimental evidence and high-certainty affirmative statements is likely to assist organism database curators, by pointing at the text that is most likely to contain evidence relevant to their curation efforts. On the other hand, identifying negative and weakly supported statements is useful to researchers who look to identify new research directions and unresolved issues. We thus expect that the accurate automated text categorization along the suggested dimensions will support a wide variety of biomedical applications.

ACKNOWLEDGEMENTS

We are grateful to the group of eight dedicated annotators for their efforts in producing the corpus.

Funding: This work is supported by H.S.'s NSERC Discovery Grant 298292-04 and CFI New Opportunities Award 10437, and by A.R.'s NSF grant (supplement to EIA-0121687).

Conflict of Interest: none declared.

REFERENCES

Blaschke,C. *et al.* (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology (ISMB)*.

Boutell,M.R. *et al.* (2004) Learning multi-label scene classification. *Pattern Recognit.*, **37**, 1757–1771.

Chang,C and Lin,C. (2001) LIBSVM: a library for support vector machines.Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (last accessed date October, 2006).

Craven,M. and Kumlien,J. (1999) Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology (ISMB)*.

Cohen,K.B. and Hunter,L. (2005) Natural language processing and systems biology. Springer series on computational biology. In Dubitzky,W. and Azuaje,F. (eds) *AI and Systems Biology*. Springer, pp. 147–174.

Friedman,C. *et al.* (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17**, S74–S82.

Ghamrawi,N. and McCallum,A. (2005) Collective multi-label classification. In *Proceedings of the Conference on Information and Knowledge Management* (*CIKM*).

Hastie,T. *et al.* (2003) *The Elements of Statistical Learning.* Springer.

Hersh,R.T. *et al.* (2006) TREC 2005 genomics track overview. In *Proceedings of the Text Retrieval Conference, (TREC'05), National Institute of Standards and Technology (NIST)*.

Krallinger,M. and Valencia,A. (2005) Text-mining and information-retrieval services for molecular biology. *Genome Biol.*, **6**, 224.

Krauthammer,M. *et al.* (2002) Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics*, **18** (Suppl. 1): S249–S257.

Kudo,T. and Matsumoto,Y. (2000) Use of support vector learning for chunk identification. In *Proceedings of the 4th Conference on CoNLL-2000 and LLL-2000*, pp. 142–144.

Langer,H. *et al.* (2004) Text type structure and logical document structure. In *Proceedings of the ACL Workshop on Discourse Annotation*.

Light,M. *et al.* (2004) The language of bioscience: facts, speculations, and statements in between. In *HLT-NAACL: BioLink'04*. pp. 17–24.

Medlock,B. and Briscoe,T. (2007) Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the meeting of the Association for Computational Linguistics* (*ACL*).

Mizuta,Y. and Collier,N. (2004) Zone identification in biology articles as a basis for information extraction. In Collier, N. (ed.) *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA)*.

Nigam,K. *et al.* (1999) Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*. pp.61–67.

Pan,F. (2006) Multi-dimensional fragment classification in biomedical text.MSc Thesis, School of Computing, Queen's University, Kingston, Ontario. Available at http://www.cs.queensu.ca/home/shatkay/papers/FengxiaPanTheis.pdf (last accessed date August, 2008).

Porter,M (1980) An algorithm for suffix stripping. *Program*, **14**, 130–137.

Raychaudhuri,S. *et al*. (2003) Computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Res*., **31**, 4553–4560.

Shatkay,H. (2005) Hairpins in bookstacks: information retrieval from biomedical text. *Brief. Bioinform.*, **6**, 222–238.

Shatkay,H. *et al.* (2000) Genes, themes and microarrays: using information retrieval for large scale gene analysis. *Proceedings of International Conference on Intelligent Systems for Molecular Biology (ISMB)*. pp. 317–328.

Smith,L. *et al.* (2004) MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics*, **20**, 2320–2321.

Tanabe,L. and Wilbur,W.J. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics*. **18**, 1124–1132.

Teufel,S. *et al*. (1999) An annotation scheme for discourse-level argumentation in research articles. *Proceedings of EACL.*

Van Rijcsbergen,C. (1979) *Information Retrieval*. 2nd edn. Butterworhs, London, UK.

Wilbur,W.J. *et al*. (2006) New directions in biomedical text annotations: definitions, guidelines and corpus construction. *BMC Bioinformatics*, **7**, 356.

Yang,Y. and Pederson,J. (1997) A comparative study on feature selection in text categorization. In *Proceedings of the International Conference on Machine Learning, (ICML)*.

Yeh,A.S. *et al*. (2003) Evaluation of text data mining for database curation: lessons learned from the KDD challenge cup. *Bioinformatics*, **19**, i331–i339.