William Hersh, MD
is Professor and Chair in the Department of Medical Informatics & Clinical Epidemiology at Oregon Health & Science University. He is well known for his work in information retrieval, particularly in health and biomedical contexts. He currently serves as Chair of the TREC Genomics Track. He has published over 100 scientific papers and is author of the book 'Information Retrieval: A Health and Biomedical Perspective'.

Keywords: *biomedical text mining, information retrieval, evaluation, recall, precision, Text Retrieval Conference*

Evaluation of biomedical text-mining systems: Lessons learned from information retrieval

William Hersh

Date received (in revised form): 8th June 2005

Abstract

Biomedical text-mining systems have great promise for improving the efficiency and productivity of biomedical researchers. However, such systems are still not in routine use. One impediment to their development is the lack of systematic and rigorous evaluation, comparable to the approaches developed for information retrieval systems. The developers of text-mining systems need to improve both test collections for system-oriented evaluation and undertake user-oriented evaluations to determine the most effective use of their systems for their intended audience.

INTRODUCTION

The goal of text-mining systems is to help biomedical researchers extract knowledge from the biomedical literature to facilitate new discovery in a more efficient manner.[1,2] The other papers in this special issue describe the motivations, applications and results of specific text-mining systems and algorithms. While most of these approaches have been assessed with evaluations showing how they perform and where they falter, these assessments are still very limited in discerning the larger role of text mining as a tool for real-world biomedical researchers.

This paper reviews the state of the science in evaluation of text-mining systems, with a particular focus on the lessons learned from similar work done in the more mature area of information retrieval (IR). We start with a broad view of research methods and technology assessment in biomedicine generally. The rest of the paper then provides an overview of the measures used in IR; a review of the largest IR evaluation initiative, the Text Retrieval Conference (TREC); and a critique of current evaluation methods in biomedical text mining.

EVALUATIVE RESEARCH IN BIOMEDICINE

The field of biomedicine has a long history of evaluative research. Certainly no one would recommend a medical treatment without its thorough evaluation, including the gold standard, the randomised controlled trial.[3] A decade-old primer on technology assessment by Littenberg is still pertinent today, noting that biomedical technologies must progress through biological plausibility, technical feasibility, intermediate outcomes (eg improved laboratory test results), patient outcomes (eg effective treatment or prevention of disease) and societal outcomes (improved cost-effectiveness of healthcare).[4]

There is no reason why information technologies in medicine should not be held to a similar standard. Indeed, Stead *et al*. outlined a framework for development of clinical informatics applications that carries out appropriate analyses at the levels of system specification, component development, combination of

William Hersh,
Department of Medical Informatics & Clinical Epidemiology,
Oregon Health & Science University,
3181 SW Sam Jackson Park Rd, BICC,
Portland, OR 97239, USA

Tel: +1 503 494 4563
Fax: +1 503 494 4551
E-mail: hersh@ohsu.edu

components into a system, integrating the system into an environment, and its routine use.[5] There are different studies that are most appropriate for these different levels of system development. Hersh and Hickam authored an overview of how well clinicians use IR systems, categorising the questions that research should ask, such as whether systems actually get used, how well they retrieve relevant documents, and ultimately whether they help the user achieve their intended task, such as answering a clinical question.[6] They also developed the OHSUMED test collection, which provides a realistic subset of MedLine, over 100 real-user topics, and comprehensive relevance judgments.[7,8]

Unlike IR systems, text-mining systems are not in routine use now

Owing to a paucity of research, there is no comparable compilation of how biomedical researchers use IR systems. The Hersh and Hickam review, along with additional related research,[9,10] does provide some insights for biomedical IR generally. For example, the type of indexing (human-assigned Medical Subject Headings [MeSH] *v.* automatically extracted key words) and retrieval (Boolean *v.* natural language) do not appear to have substantial impact on retrieval results. In addition, the variance across users appears to be much more substantial than across systems, so the more substantive way to improve usefulness of IR systems is probably more at the user than the system level.

The role of IR systems in the text-mining pipeline is underappreciated

As we will see in this paper, IR systems are much further along than text-mining systems both in their amount of routine use (ie probably all real-world biomedical researchers have used PubMed and Google, whereas very few have used a text-mining system) and in the sophistication of their evaluation. This does not mean that text-mining systems will not someday have a substantial role as a tool for such researchers. At the present time, however, despite pronouncements they are ready for widespread adoption,[11] biomedical text-mining systems are not in routine use.

Not only must text-mining researchers appreciate the experience of IR system evaluation, they also need to better understand the role that IR systems will play in text mining. To discern the difference between IR and text-mining systems, Figure 1 shows this author's 'funnel of knowledge-based information', which demonstrates that IR systems play a fundamental role in the pipeline of providing the appropriate amount of literature to text-mining algorithms for their effective operation. Indeed, most of the text-mining evaluations identified by this review use corpora on the order of dozens or hundreds of documents. In order to winnow the number of papers down from the current 13 million documents in MedLine, well-performing IR techniques are essential.

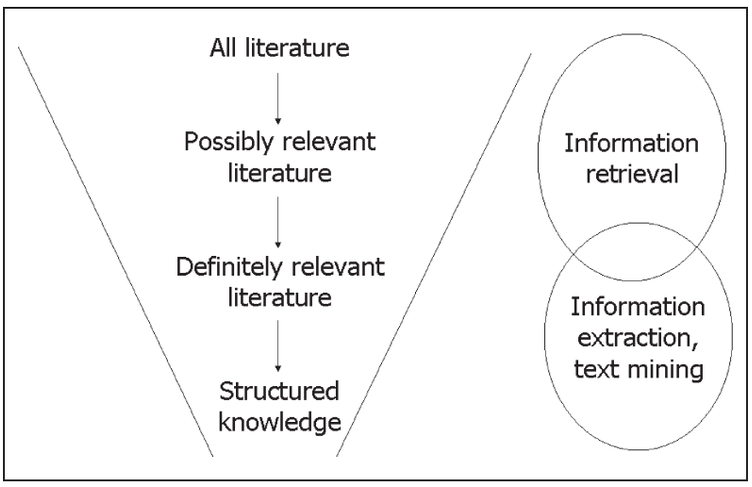

Figure 1: The 'funnel of knowledge-based information', demonstrating how the structuring and understanding of knowledge progresses from all biomedical literature to more refined and relevant amounts. At larger amounts of literature, information retrieval techniques are more predominant, while once the relevant literature has been defined, information extraction and text mining are more key

Although there are many types of research in science broadly, most technology assessments (whether biomedical treatments or information systems) use *comparative research*. The goal of comparative research is to determine whether one approach to something works better than another. It uses experiments whose purpose is to learn the truth about the world by means of the

most objective, disinterested process possible.

In comparative research, there is usually some sort of *hypothesis* being tested. Comparative research is likely to ask questions about a *population*. In biomedical research, that population may be a group of individuals at risk for a disease, perhaps because they have a gene that is known to be associated with it. Research questions with information technologies are similar. For example, researchers might want to know how well a system provides relevant information to users. Since information always has intended users, the population of study will be a group specified by one or more characteristics, which may be demographic, professional, educational or other attributes. Different users (eg clinicians, consumers, genomics researchers) are different populations with different needs and expectations for searching.

Different users have differing needs and expectations

In scientific studies, whether in medicine or IR, it is impractical to observe entire populations. As such, researchers study *samples* that represent populations. Ideally, the samples are chosen to accurately represent the populations and interventions therein being studied. The samples should be selected randomly within the population to ensure that members have an equal chance of being included in the sample. If an intervention is going to be applied, it needs to be certain that any member of the population being randomised can end up in any study group.

Results that are not statistically significant may be due to low statistical power

After the intervention has been studied and results obtained, the next step is to determine whether the results are truly correct, ie the sample reflects the truth of the larger population. Researchers hope that their results reflect the truth; however, there are other possibilities. There are two general categories of error in extrapolating the observations of a sample to the entire population, bias and chance.

Error in research arises from bias or chance

Bias is the systematic error introduced into research studies that results from improper sampling, measurements or other problems. Many authors have described the different types of bias that can influence the results of experiments.[12] These include the use of experimental methods or test data that, perhaps inadvertent to the researcher, give one approach better results over another.

The other category of error is *chance*. In this type of error, an observation occurs because the randomisation process has inadvertently produced a sample that does not accurately represent the population. One method that helps to minimise chance error is the use of statistical analysis. The appropriate statistical test helps determine how unlikely the results are due to chance. In fact, the major purpose of *inferential statistics* is to determine how well observations of the sample can be extrapolated to the whole population. These statistics identify two types of error, commonly called alpha and beta. *Alpha* error determines how likely it is that a difference in treatment groups is actually due to chance. This is the famous 'p value', with a level of $p < 0.05$ indicating there is only a 1 in 20 chance that an observation is due to random error. Note that this does not say the result is not due to random error, only that such an error is highly unlikely. *Beta* error states how likely it is that a non-difference in groups truly represents no difference. This type of error is often overlooked in scientific research, in that results found to be 'not statistically significant' may very well have a low likelihood of being able to detect a difference in the first place.[13] The measure of a study to avoid beta error is called its *statistical power*.

Within the comparative evaluation of systems, additional distinctions can be made. One distinction is to denote whether the evaluation focuses on a whole system (called *macroevaluation*) or some or all of its specific components (*microevaluation*).[14] Another distinction is to note whether evaluations are *system-oriented*, where the focus is on the system and/or its components, or *user-oriented*,

where the focus of the evaluation how systems fare in the real world.

It should be noted that there are other types of research besides comparative research. There is also qualitative or subjectivist research,[12] the results of which can provide additional insight into scientific observations that comparative measurements may not. Some of the techniques in this type of research include *ethnography*, where investigators immerse themselves in the environment in which the system is being used, and *focus groups*, where selected panels participate in a structured interview process and attempt to obtain a consensus on something. Significant effort is devoted to analysing the narrative of participants to identify patterns and insights. A related type of research is *usability testing*, in which users are provided tasks to perform with the system and their actions are captured and analysed.[15]

System-oriented IR evaluation most frequently employs recall and precision

INFORMATION RETRIEVAL (IR) EVALUATION

The goal of most system-oriented research in retrieval or mining systems is to compare approaches based on their appropriate retrieval of objects (documents, concepts, etc) that should be retrieved and the non-retrieval of those that should not. It turns out that many fields quite unrelated to IR and text mining use these sorts of measure. One of the most notable of these is that of diagnostic test evaluation. Table 1 is a two-by-two contingency table that reflects the possible outcomes when the results of a system are compared with some sort of gold standard. Relative to that gold standard, retrieved items can be true positive (TP), false positive (FP), false negative (FN) and true negative (TN).

Table 1: The two-by-two contingency table for defining the measures of how well a test (eg information retrieval or text-mining system or medical diagnostic test) compares with a gold standard

		Gold standard		
		Positive	**Negative**	**Total**
Test results	Positive	TP	FP	TP + FP
	Negative	FN	TN	FP + TN
	Total	TP + FN	FP + TN	TP + FP + FN + TN

In the case of IR, the gold standard is a relevant document and the test result is a retrieved document. The total number of relevant documents is TP + FN, while the total number of retrieved documents TP + FP. The number of documents that are retrieved and relevant is TP. The most commonly used measures in IR are recall and precision. These are sometimes called the relevance-based measures because they measure the proportion of relevant documents retrieved from the database (recall) or from within the search (precision).

More formally, *recall* is the proportion of relevant documents retrieved from the database:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1)$$

In other words, recall measures the fraction of all relevant documents have been obtained from the database. One problem with equation 1 is that the denominator implies that the total number of relevant documents for a query is known. For all but the smallest of databases, however, it is unlikely, perhaps even impossible, for one to succeed in identifying all relevant documents in a database. Thus most studies use the measure of *relative recall*, where the denominator is redefined to represent the number of relevant documents identified by multiple searches on the topic.

Precision is the proportion of relevant documents retrieved in the search:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

Precision measures the fraction of the retrieved documents that are relevant.

Table 1 is very similar to the matrix used to calculate the medical diagnostic test performance measures of sensitivity and specificity. In fact, if 'relevance' is changed to 'presence of disease' and

'number retrieved' is changed to 'number with positive diagnostic test result', then recall is identical to *sensitivity*, while precision is the same as *positive predictive value*. Another measure commonly used is diagnostic test assessment is *specificity*, which is the proportion of items correctly not retrieved or classified:

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \quad (3)$$

Specificity is typically not used in IR research results, since the numbers of both relevant and retrieved articles for a given query tend to be small relative to the very large database. With these large databases, therefore, specificity would almost always approach 100 per cent.

The usual system-oriented approach to IR evaluation involves the use of a test collection, which consists of a set of documents, topics (sometimes called queries, although most reserve that work for the actual text entered into the system) and relevance judgments. In an experiment, sometimes called a *run* in this context, the selected retrieval measure is determined for each topic and an average is taken of all of them. A great deal has been written about 'optimal' test collections, noting that their document types and quantities, along with their topics or queries, reflect the real-world needs of intended users of systems.[16] Test collections should also have assessment of the reliability of its relevance judgments,[17] which is usually done with measures such as *kappa*.[18]

MAP is the most common aggregate measure used in recent IR evaluations

One challenge in interpreting recall and precision is that there tends to be a trade-off between the two. That is, a search that aims to achieve high recall may obtain low precision, or vice versa. As such, we might desire to combine them into a single measure. This has been done in a number of ways, although one caveat with combining the two measures is that the resulting measure loses the real-world meanings that are clear with simple recall and precision.

One approach to combining recall and precision is the *F* measure,[19] which combines recall and precision as follows:

$$F = \frac{(1 + \beta^2) \times \text{Recall} \times \text{Precision}}{(\beta^2 \times \text{Precision}) + \text{Recall}} \quad (4)$$

The variable β indicates the relative value of precision. A value of $\beta = 1$, which is usually used, indicates equal value of recall and precision, whereas lower values indicate more emphasis on precision and higher values indicate more emphasis on recall.

Another method of combining recall and precision is the recall-precision table, whose use is limited to IR systems that rank their document output.[20] In a recall-precision table, a typical approach is to use intervals of 0.1 (or 10 per cent), with a total of 11 intervals from a recall of 0.0 to 1.0. The table is built by determining the highest level of overall precision at any point in the output for a given interval of recall. Thus, for the recall interval 0.0, one would use the highest level of precision at which the recall is anywhere greater than or equal to zero and less than 0.1. Summary results are usually reported as the average of precision at each point of recall for all the topics in the run.

A related approach that has been used more frequently in recent times has been the mean average precision (MAP), which is similar to precision at points of recall but does not use fixed recall intervals or interpolation.[21] Instead, precision is measured at every point at which a relevant document is obtained, and the average precision for a topic is calculated by averaging the precision at each of these points. MAP is then calculated by taking the mean of the average precision values across all topics in the run. MAP has been found to be a stable measure for combining recall and precision, but suffers from its value arising from statistical aggregation and having no real-world meaning.

There are other approaches to comparing ranked output. One is to fix the level of documents retrieved (eg 30), which gives recall and precision points that can then be compared by means of

ranked output against those of other systems, including systems that do not use weighting at all. An additional approach further develops the medical diagnostic testing analogy described earlier. In this case, precision (or positive predictive value) is converted to specificity, generating a receiver operating characteristic (ROC) curve for each query.[22]

Relevance-based measures have limitations

There are a number of limitations with recall and precision, especially when they are sole values used to assess IR systems. A number of practical problems arise when one is attempting to measure them. For example, what is a query? Is it a single search against the system, or is it all the searches on a given topic? Does it occur in a single setting, or can the user come back to the system after utilising the documents, perhaps hours or days later? In the latter, what should define the base of interactions that allow measurement of recall and precision for a search? An additional problem lies in measuring them across a set of topics. Is it adequate to simply take the mean? It turns out that in most results using test collections, variance across queries is typically quite large. One consequence of this is that the mean values tend to obscure a great deal of variation of the individual measurements.

Another problem is deciding what constitutes a retrieved document. Should it be the document 'surrogate' (ie the title with or without other brief information) presented after a search has been entered, or it should it only be the documents examined in more detail by the searcher (ie the whole MedLine reference or Web page viewed after the surrogate has been chosen)? And again, should documents retrieved by a poor, perhaps erroneous, original search be counted as 'retrieved?'

The most important forum for IR evaluation is TREC

An additional challenge with recall and precision is what constitutes a meaningful difference across systems or algorithms. Clearly we want to build and use IR systems that retrieve relevant documents. But what level of recall or precision will result in a more meaningful or productive experience for the real user? There is some evidence to suggest that the differences in recall and precision obtained by typical state-of-the-art retrieval do not affect performance on real-world tasks, such as finding instances of an information need or answering questions.[9,10] The likely reason for the lack of a difference is that the differences among users (eg their knowledge, underlying skills and cognitive abilities) overwhelm any differences in the number of relevant documents retrieved.[23]

Things get even murkier when we start to think about relevance. So far, we have defined relevance as a document meeting an information need that prompted a query. This fixed view of relevance makes recall and precision very straightforward to calculate. But as it turns out, relevance is not quite so objective. For example, relevance as judged by physicians has a moderately high degree of variation.[8,10,24] In each of these studies, the kappa statistic showed only a 'fair' level of agreement of relevance judges. Schamber *et al*. have noted there are really two types of relevance: a system-oriented topical view and a user-oriented situational view.[25] The former is helpful in evaluations that use test collections, but the latter is more reflective of real-world searchers.

THE TEXT RETRIEVAL CONFERENCE

No overview of IR evaluation can ignore TREC, the Text REtrieval Conference[26] organised by the US National Institute for Standards and Technology (NIST[27]).[21] Started in 1992, TREC has provided a test-bed for evaluation and a forum for presentation of results. TREC is organised as an annual event at which the tasks are specified, and queries and documents are provided to participants. Participating groups submit 'runs' of their systems to NIST, which calculates the appropriate performance measure, usually recall and precision.

One of the motivations for starting TREC was the observation that much IR evaluation research (prior to the early

TREC now has many tracks based on categories of IR research interest

1990s) was done on small test collections that were not representative of real-world databases. Furthermore, some companies had developed their own large databases for evaluation but were unwilling to share them with other researchers. TREC was therefore designed to serve as a means to increase communication among academic, industrial and governmental IR researchers. Although the results were presented in a way that allowed comparison of different systems, conference organisers advocated that the forum not be a 'competition' but instead a means to share ideas and techniques for successful IR. In fact, participants are required to sign an agreement not to use results of the conference in advertisements and other public materials.[21]

The original TREC conference featured two common tasks for all participants. An *ad hoc retrieval* task simulated an IR system, where a static set of documents was searched using new topics, similar to the way a user might search a database or Web search engine for the first time. A *routeing* task, on the other hand, simulated a standing query against an oncoming new stream of documents, similar to a topic expert's attempt to extract new information about his or her area of interest. The original tasks used newswire and government documents, with queries created by US government information analysts. Relevance judgments were performed by the same information analysts who created the queries.[28]

The main findings from the early TREC conferences were that modest benefits in retrieval performance could be attained consistently by several techniques:[21]

- Document weighting – the Okapi and pivoted normalisation approaches.
- Query expansion – expanding queries with terms from high-ranking documents in the initial retrieval.
- Passage retrieval – giving higher weight to documents having passages with high concentrations of query terms.

Each of these approaches improved measures such as MAP by 5–15 per cent, giving slightly higher results when used in combination.

By the third TREC conference (TREC-3), interest was developing in other IR areas besides ad hoc searching and routeing. At that time, the conference began to introduce tracks geared to specific interests. In an overview of TREC, Voorhees recently categorised the tasks and tracks associated with them:[29]

- Static text – Ad Hoc.
- Streamed text – Routeing, Filtering.
- Human in the loop – Interactive.
- Beyond English (cross-lingual) – Spanish, Chinese and others.
- Beyond text – OCR, Speech, Video.
- Web searching – Very Large Corpus, Web.
- Answers, not documents – Question-Answering.
- Retrieval in a domain – Genomics.

In some TREC tracks, different evaluation measures have been necessary because the notion of a relevant document no longer is solely how performance is appropriately measured. In the Question-Answering track, for example, the goal of the task is to provide an answer within a small span of the document.[30] Furthermore, only the top-ranking answer string that contains the answer is important, since any further answers would be redundant. This track therefore uses the mean reciprocal rank (MRR), which is calculated based on the mean of the reciprocal rank (RR) of the answer string nominated by the system:

$$\mathrm{RR} = \frac{1}{\text{Rank of string with answer}} \quad (5)$$

The Filtering track was an outgrowth of the routeing task, with the difference that the user is truly interested in a stream of documents as opposed to a ranked list, which was how the original routeing task was evaluated.[31] The information filtering setting is typically in a busy work environment, where the user is confronted by a steady stream of new documents and wants to minimise the time spent reading non-relevant documents. This track's evaluation measure therefore is based on a utility score that includes a penalty for non-relevant documents that are retrieved:

While mostly focusing on system-oriented evaluation, TREC has also looked at user-oriented evaluation

$$\text{Utility} = (u_r \times \text{relevant documents retrieved}) + (u_{nr} \times \text{non-relevant documents retrieved}) \quad (6)$$

where u_r and u_{nr} are relative utilities of the value of retrieving relevant and non-relevant documents, respectively. In recent iterations of the track, the values of u_r and u_{nr} have been set at 2 and -1.[32] The only part of TREC to operate in the biomedical domain, prior to the TREC Genomics Track, was the use of the OHSUMED test collection by the Filtering track. The task was to assess the ability of systems to 'classify' MEDLINE references into MeSH categories in this track for one year.[33]

The Genomics Track is the first domain-specific track in TREC

The TREC community was not quick to recognise the importance of Web searching. This was later rectified by the addition of a Web track, but the techniques that have been found to be most valuable on the Web (eg the Google PageRank algorithm[34]) did not emanate from the academic IR community. Of course, it should be noted that the PageRank approach does not necessarily work for all types of IR. Not only does some IR content lack the links that give PageRank its power, but it is not clear that ranking documents by 'majority vote' is appropriate for content such as the biomedical literature.

A small but important thread of activity at TREC has been interactive retrieval evaluation. For several years there was an Interactive track. The goal of this track was to determine whether the systems and algorithms of the other tracks, in the hands of real users, achieved the results shown by system-oriented studies. Although the track never achieved widespread participation, most of its results did show that one could not assume *a priori* that techniques found effective by system-oriented evaluation would automatically benefit real users.[35]

The TREC experiments have also led to research about evaluation itself. Voorhees, for example, has assessed the impact of different relevance judgments on results in the TREC ad hoc task.[28] In the TREC-6 ad hoc task, over 13,000 documents among the 50 queries had duplicate judgments. Substituting one set of judgments for the other was found to cause minor changes in MAP for different systems but not their order relative to other systems. In other words, different judgments changed the MAP number but not the relative performance among different systems. Zobel has demonstrated that the number of relevant documents in the ad hoc track is likely underestimated, hence recall may be overstated.[36]

Although TREC has primarily focused on non-biomedical newswire and government documents, it has taken on 'retrieval in a domain', in particular the genomics domain, in recent years. Led by this author, the TREC Genomics Track has aimed to assess IR in the biomedical domain. Owing to the community's interest in information extraction and text mining as well, other tasks besides pure ad hoc retrieval have been developed. In the most recent year of the track (2004), one task assessed ad hoc retrieval from topics captured from real biologists, while another assessed categorisation decisions made by curators of Mouse Genome Informatics (MGI[37]) system.[38]

Participants in the 2004 ad hoc retrieval

task obtained a wide range of results. In general, they achieved benefit with both domain-specific techniques, such as expanding queries with gene name and other synonyms for terms, and non-domain-specific techniques shown previously to work in TREC generally, including more advanced document weighting as well as query expansion. Attempts at recognising the names of specific genes or other entities for adding synonyms or other information were in general less successful.

The categorisation task had two subtasks. One was a document 'triage' task, akin to filtering, where the system had to decide whether to designate the article as having experimental data to warrant assignment of terms from the Gene Ontology.[39] As such, the utility measure in equation 6 was employed to evaluate the task, although the parameters were set differently, reflecting the importance to MGI of not missing a relevant document over designated a non-relevant article for analysis. As such, the values of u_r and u_{nr} were set at 20 and −1 respectively. Although many groups tried elaborate machine learning-based approaches, none improved over very simple ones. The other subtask focused on the categorisation of documents as having terms assigned from zero to three of the Gene Ontology hierarchies.

BIOMEDICAL TEXT-MINING EVALUATION

With the understanding of the context of text mining and the experience of IR research, we can turn our attention to the evaluation of text-mining systems. An exhaustive review of all evaluations done in text mining cannot be given, especially since many of them were discussed in a recent issue of this journal.[2] Therefore the goal of this section is to critique the current state of text-mining evaluation and suggest directions for future research.

Much text-mining evaluation still focuses on system-oriented microevaluations

Most research in text mining still focuses on the development of specific functions or algorithms, usually evaluated by system-oriented microevaluations. A TREC-like forerunner in the area of information extraction, itself a conceptual predecessor of text mining, was the Message Understanding Conference (MUC[40]). Most of MUC was devoted to 'template filling' from processing of documents.[41] Systems that were able to recognise 'named entities', usually through use of natural language processing techniques, performed best.

While there are some 'complete' text-mining systems, eg Textpresso[42] and MedScan,[43] none is really in routine use by end-users. As noted above, this is contrast to IR systems, where almost all biomedical researchers use applications such as PubMed and Google. From the vantage point of Stead *et al.*,[5] whereas IR systems function at the level of routine use, text-mining research is still concerned with individual components. Because of this, most evaluation is still system-oriented, or perhaps more appropriately described, algorithm-oriented. That is, the focus of the evaluation has been on the specific function of the algorithm, such as finding the presence of a gene name or its normalised term, identifying synonyms or acronyms, or the automated annotation by a vocabulary, such as the Gene Ontology.[39]

In assessing the landscape of evaluative research in text mining, it is clear there are deficiencies in test collections. This stems in part from the microevaluation view that most evaluations take. Researchers have chosen, perhaps rightly so given the state of the field, to focus on the evaluation of specific components of systems, such as named entity recognition, classification of documents, or detection of relationships. As with retrieving relevant documents, such functions are important, but focusing on them individually instead of in a complete system may mask interactions between them that are essential for the functionality for a real user.

Another problem with text-mining test collections is that different researchers not only use different collections, but also use

those that do exist in different ways. A classic example of this is in named entity recognition, probably the most studied (or at least most published) aspect of biomedical text mining. The varying results one obtains when differing assumptions are made as to whether full or partial matching is considered the gold standard, or full collections or partial collections are used, is exemplified by the results of Hou and Chen.[44]

A number of researchers have developed their own collections for use by their own research systems. For example, the systems mentioned above, Textpresso and MedScan, have their own test collections that have been used only for their evaluation. Similarly, the well-known Abgene tagger has a test collection that has not been re-used by other research groups.[45]

There are few large-scale comprehensive test collections for text-mining evaluation

So one challenge for the text-mining community is to standardise, akin to TREC, on one or a small number of high-quality test collections and measures by which systems are assessed. There have been some test collections that have had some amount of reuse. One such collection is GENIA.[46] Focused on named entity recognition, GENIA has been used by many researchers, although not always in a consistent way. It was also used in the recent shared named entity recognition task of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA).[47] GENIA consists of 2,000 MedLine abstracts on the topic of human blood cell transcription factors and has been marked up for both part of speech and several dozen biological entities. The human inter-tagger reliability of GENIA's tagging has not been assessed.

There are also two test collections developed by researchers from MITRE who have organised challenge evaluations around them, the Knowledge Discovery from Databases (KDD) Challenge Cup[48] and the Critical Assessment for Information Extraction in Biology (BioCreAtIvE).[49] The latter represents what is probably the most comprehensive effort to date in developing a test collection and challenge evaluation.

BioCreAtIvE consisted of two general tasks, each with subtasks. The first task focused on named entity recognition, with Task 1A involving recognition of phrases in the text representing gene and/or protein names,[50] and Task 1B involving the normalisation of gene identifiers with an article abstract.[51] Task 1A used 7,500 sentences from MedLine abstracts for training and 2,500 sentences for testing. Task 1B used a collection of training and test MedLine records from the fly, mouse and yeast organisms for the recognition of normalised names of genes. The second task of BioCreAtIvE focused on automated annotation of proteins using Gene Ontology terms from the full text of articles of the Biomed Central text-mining corpus.[52] Unlike other text-mining test collections, some analysis was done in BioCreAtIvE assessing inter-annotator agreement and its implications in the results.[53]

Most of the rest of text-mining research has focused on test collections developed by individual research groups. These collections are used for a variety of text-mining purposes. While they are described in the literature, they are not always available to other groups. For example, at least one of these collections, which will not be mentioned by name here, is not available to other research groups because of copyright issues. A major problem with collections that cannot be shared is that the results of reported findings cannot be replicated or improved upon.

While some of the developments with test collections are encouraging, better collections are still needed. Most of the collections used to date are extremely small, in the order of hundreds or at most thousands of documents. While the resources to build such collections are not trivial, the large amount of funding going into bioinformatics and computational biology would seem to justify the development of large and realistic test

collections. However, further test collections should not just mimic what has been done to date, but instead focus on tasks likely to affect real users. Perhaps another thread of research should focus on better defining the information tasks of biomedical researchers, who have been vastly less studied than other information users, although there are some notable exceptions.[54–56]

Another challenge to the biomedical text-mining community is to take more seriously the IR functionality at the front end of their systems. Most research systems have operated under an implicit assumption that they will start with the 'right' documents. However, IR researchers can point out that honing a collection of documents that are only or mostly relevant is not a trivial task. The TREC Genomics Track, organised by this author, will certainly continue to promote this view.

In addition to being cognisant of IR, the text-mining community must also expand its horizons and begin thinking beyond system-oriented microevaluations. Even if the components of systems have not been perfected, the community must begin to fathom how these tools will be brought to bear on the real information-related problems of biomedical researchers. There is still a great deal of real-world use and evaluative research of operational IR systems. Furthermore, some of this research demonstrates that what is measured in system-oriented evaluations is not necessarily what translates into a user's ability to better perform information tasks.[9,10] This means that text-mining research must move beyond name-finding and document-ranking algorithms to develop tools that help real users improve the quality and efficiency of their biomedical research.

Future text-mining research must focus on real uses of systems by real users

CONCLUSIONS

If the promise of text mining to enhance biomedical research is to be met, better evaluation is essential. This will not only help the field better determine what approaches work best, but also provide insight into how systems can best enhance the work of their intended users. The field must follow the lead of IR researchers to not only develop better system-oriented research resources, such as test collections, but also begin to focus on user-oriented evaluations to determine how systems will be most effective in real-world settings.

Acknowledgments

This work was supported by grant ITR-0325160 from the National Science Foundation and earlier grants from the National Library of Medicine.

References

1. Hirschman, L., Park, J. C., Tsuhii, J. *et al.* (2002), 'Accomplishments and challenges in literature data mining for biology', *Bioinformatics*, Vol. 18, pp. 1553–1561.
2. Cohen, A. M. and Hersh, W. R. (2005), 'A survey of current work in biomedical text mining', *Brief. Bioinformatics*, Vol. 6, pp. 57–71.
3. Sackett, D. L., Straus, S. E., Richardson, W. S. *et al.* (2000), 'Evidence-Based Medicine: How to Practice and Teach EBM', Churchill Livingstone, New York.
4. Littenberg, B. (1992), 'Technology assessment in medicine', *Acad. Med.*, Vol. 67, pp. 424–428.
5. Stead, W. W., Haynes, R. B., Fuller, S. *et al.* (1994), 'Designing medical informatics research and library-resource projects to increase what is learned', *J. Amer. Med. Informatics Assoc.*, Vol. 1, pp. 28–33.
6. Hersh, W. R. and Hickam, D. H. (1998), 'How well do physicians use electronic information retrieval systems? A framework for investigation and review of the literature', *J. Amer. Med. Assoc.*, Vol. 280, pp. 1347–1352.
7. Hersh, W. R. and Hickam, D. H. (1994), 'The use of a multi-application computer workstation in a clinical setting', *Bull. Med. Library Assoc.*, Vol. 82, pp. 382–389.
8. Hersh, W. R., Buckley, C., Loene, T. J. and Hicham, D. (1994), 'OHSUMED: An interactive retrieval evaluation and new large test collection for research', in 'Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Dublin, Ireland, Springer, New York, pp. 192–201.
9. Hersh, W., Turpin, A., Price, S. *et al.* (2001), 'Challenging conventional assumptions of automated information retrieval with real

users: Boolean searching and batch retrieval evaluations', *Inform. Proc. Manage.*, Vol. 37, pp. 383–402.

10. Hersh, W. R., Crabtree, K., Hickam, D. H. *et al.* (2002), 'Factors associated with success for searching MEDLINE and applying evidence to answer clinical questions', *J. Amer. Med. Informatics Assoc.*, Vol. 9, pp. 283–293.
11. Rebholz-Schuhmann, D., Kirsch, H. and Couto, F. (2005), 'Facts from text – is text mining ready to deliver?', *PLoS Biol.*, Vol. 3, p. e65.
12. Friedman, C. P. and Wyatt, J. C. (1997), 'Evaluation Methods in Medical Informatics', Springer, New York.
13. Halpern, S. D., Karlawish, J. H. T. and Berlin, J. A. (2002), 'The continuing unethical conduct of underpowered clinical trials', *J. Amer. Med. Assoc.*, Vol. 288, pp. 358–362.
14. Lancaster, F. W. and Warner, A. J. (1993), 'Information Retrieval Today', Information Resources Press, Arlington, VA.
15. Nielsen, J. (1993), 'Usability Engineering', Academic Press, San Diego, CA.
16. Sparck-Jones, K. (1981), 'Information Retrieval Experiment', Butterworth, London.
17. Saracevic, T. (1975), 'Relevance: A review of and a framework for the thinking on the notion in information science', *J. Amer. Soc. Information Sci.*, Vol. 26, pp. 321–343.
18. Light, R. J. (1971), 'Measures of response agreement for qualitative data: Some generalizations and alternatives', *Psychol. Bull.*, Vol. 76, pp. 365–377.
19. vanRijsbergen, C. J. (1979), 'Information Retrieval', Butterworth, London.
20. Salton, G. (1983), 'Introduction to Modern Information Retrieval', McGraw-Hill, New York.
21. Voorhees, E. M. and Harman, D. (2000), 'Overview of the Sixth Text REtrieval Conference (TREC)', *Information Proc. Manage.*, Vol. 36, pp. 3–36.
22. Hanley, J. A. and McNeil, B. J. (1983), 'A method of comparing the areas under receiver operating characteristic curves derived from the same cases', *Radiology*, Vol. 148, pp. 839–843.
23. Turpin, A. H. and Hersh, W. (2001), 'Why batch and user evaluations do not give the same results', in 'Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', New Orleans, LA, ACM Press, New York, pp. 225–231.
24. Haynes, R. B., McKibbon, K. A., Walker, C. J. *et al.* (1990), 'Online access to MEDLINE in clinical settings', *Ann. Intern. Med.*, Vol. 112, pp. 78–84.
25. Schamber, L., Eisenberg, M. B. and Nilan, M. S. (1990), 'A re-examination of relevance: Toward a dynamic, situational definition', *Information Proc. Manage.*, Vol. 26, pp. 755–776.
26. URL: http://trec.nist.gov
27. URL: http://www.nist.gov
28. Voorhees, E. M. (1998), 'Variations in relevance judgments and the measurement of retrieval effectiveness', in 'Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Melbourne, Australia, ACM Press, New York, pp. 315–323.
29. Voorhees, E. M. (2003), 'Overview of TREC 2003. The Twelfth Text REtrieval Conference – TREC 2003', Gaithersburg, MD, National Institute for Standards & Technology, pp. 1–13 (URL: http://trec.nist.gov/pubs/trec12/papers/OVERVIEW.12.pdf).
30. Voorhees, E. M. and Tice, D. M. (2000), 'Building a question answering test collection', in 'Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Athens, Greece, ACM Press, New York, pp. 200–207.
31. Lewis, D. D. (1995), 'Evaluating and optimizing autonomous text classification systems', in 'Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Seattle, WA, ACM Press, New York, pp. 246–254.
32. Robertson, S. and Soboroff, I. (2001), 'The TREC 2001 Filtering Track report', in 'The Tenth Text REtrieval Conference (TREC 2001)', Gaithersburg, MD, National Institute of Standards and Technology, pp. 26–37 (URL: http://trec.nist.gov/pubs/trec10/papers/filtering2_track.pdf).
33. Robertson, S. and Hull, D. A. (2000), 'The TREC-9 Filtering Track final report', in 'The Ninth Text REtrieval Conference (TREC-9)', Gaithersburg, MD, National Institute of Standards and Technology, pp. 25–40.
34. Brin, S. and Page, L. (1998), 'The anatomy of a large-scale hypertextual Web search engine', *Computer Networks*, Vol. 30, pp. 107–117.
35. Hersh, W. R. (2001), 'Interactivity at the Text Retrieval Conference (TREC)', *Information Proc. Manage.*, Vol. 37, pp. 365–366.
36. Zobel, J. (1998), 'How reliable are the results of large-scale information retrieval experiments?', in 'Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Melbourne, Australia, ACM Press, New York, pp. 307–314.
37. URL: http://www.informatics.jax.org

38. Hersh, W., Bhuptiraju, R. T., Ross, L. *et al.* (2004), 'TREC 2004 genomics track overview', in 'The Thirteenth Text Retrieval Conference (TREC 2004)', Gaithersburg, MD, National Institute of Standards and Technology (in press) (URL: http://trec.nist.gov/pubs/trec13/papers/GEO.OVERVIEW.pdf).

39. Anonymous (2004), 'The Gene Ontology (GO) database and informatics resource', *Nucleic Acids Res.*, Vol. 32, pp. D258–261.

40. URL: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

41. Chinchor, N. A. (1998), 'Overview of MUC-7/MET-2', in 'MUC-7 Proceedings', Gaithersburg, MD, National Institute of Standards and Technology (URL: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/overview.html).

42. Müller, H. M., Kenny, E. E. and Sternberg, P. W. (2004), 'Textpresso: An ontology-based information retrieval and extraction system for biological literature', *PLoS Biol.*, Vol. 2, p. e309.

43. Daraselia, N., Yuryev, A., Egorov, S. *et al.* (2004), 'Extracting human protein interactions from MEDLINE using a full-sentence parser', *Bioinformatics*, Vol. 20, pp. 604–611.

44. Hou, W. J. and Chen, H. H. (2004), 'Enhancing performance of protein and gene name recognizers with filtering and integration strategies', *J. Biomed. Informatics*, Vol. 37, pp. 448–460.

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

46. Kim, J. D., Ohta, T., Tateisi, Y. and Tsujii, J. (2003), 'GENIA corpus – semantically annotated corpus for bio-textmining', *Bioinformatics*, Vol. 19, pp. i180–182.

47. Collier, N., Ruch, P. and Nazarenko, A., eds. (2004), 'Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications', University of Geneva, Geneva, Switzerland.

48. Yeh, A. S., Hirschman, L. and Morgan, A. A. (2003), 'Evaluation of text data mining for database curation: Lessons learned from the KDD Challenge Cup', *Bioinformatics*, Vol. 19, pp. I331–339.

49. Hirschman, L., Yeh, A., Blaschke, C. and Valencia, A. (2005), 'Overview of BioCreAtIvE: Critical assessment of information extraction for biology', *BMC Bioinformatics*, Vol. 6, p. S1 (URL: http://www.biomedcentral.com/1471–2105/6/S1/S1/).

50. Yeh, A., Morgan, A., Colosimo, M. and Hirschman, L. (2005), 'BioCreAtIvE Task 1A: Gene mention finding evaluation', *BMC Bioinformatics*, Vol. 6, p. S2 (URL: http://www.biomedcentral.com/1471–2105/6/S1/S2/).

51. Hirschman, L., Colosimo, M., Morgan, A. and Yeh, A. (2005), 'Overview of BioCreAtIvE task 1B: Normalized gene lists', *BMC Bioinformatics*, Vol. 6, p. S11 (URL: http://www.biomedcentral.com/1471–2105/6/S1/S11/).

52. URL: http://www.biomedcentral.com/info/about/datamining/

53. Colosimo, M. E., Morgan, A. A., Yeh, A. *et al.* (2005), 'Data preparation and interannotator agreement: BioCreAtIvE Task 1B', *BMC Bioinformatics*, Vol. 6, p. S12 (URL: http://www.biomedcentral.com/1471–2105/6/S1/S12/).

54. Stevens, R., Goble, C., Baker, P. and Brass, A. (2001), 'A classification of tasks in bioinformatics', *Bioinformatics*, Vol. 17, pp. 180–188.

55. Tran, D., Dubay, C., Gorman, P. and Hersh, W. (2004), 'Applying task analysis to describe and facilitate bioinformatics tasks', in 'MEDINFO 2004 – Proceedings of the Eleventh World Congress on Medical Informatics', San Francisco, CA, IOS Press, Amsterdam, pp. 818–822.

56. Bartlett, J. C. and Toms, E. G. (2005), 'Developing a protocol for bioinformatics analysis: An integrated information and behavior task analysis approach', *J. Amer. Soc. Information Sci. Technol.*, Vol. 56, pp. 469–482.