# Searching for High-Utility Text in the Biomedical Literature
## A Preliminary Report

**Hagit Shatkay**
School of Computing
Queen's University
Kingston, Ontario, Canada
shatkay@cs.queensu.ca

**Andrey Rzhetsky**
Dept. of Biomedical Informatics
Columbia University
New York, New York
ar345@columbia.edu

**W. John Wilbur**
Computational Biology Branch
NCBI/NLM/NIH
Bethesda, Maryland
Wilbur@ncbi.nlm.nih.gov

**Abstract**

Much current research is concerned with extracting biomedical facts from text, so far with relatively modest results. Our work is motivated by the idea that text mining can be improved, if the system could first identify text regions that are rich in scientific content, retrieve documents that have many such regions, and focus on fact extraction from these regions. We call these parts of the text "high utility regions". In this preliminary report we describe the basic ideas, the initial steps we took, and the annotation guidelines we devised to construct a comprehensive training and test corpus that would enable us to apply and evaluate machine learning methods for identifying high-utility regions.

## Introduction

The past few years have seen an impressive growth in the amount of research dedicated to biomedical text mining, (several recent reviews include [3,4,16]). The field that originally focused on medical text [1,6,15,18] has expanded since the onset of the "genomic era" into the biological domain. Research in the area includes work on information extraction from the biomedical literature [2,7,12,19,20], as well as on information retrieval and text categorization [5,8,9,17].

The efforts on information extraction concentrate on identifying bio-entities (mostly genes and proteins) and the 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. [9]

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. Krauthammer *et al.*[10] suggested a critical examination of literature contents in molecular biology, and recent work by Light *et al*.[13] also examined the validity and reliability of statements made in the literature.

Taking the idea of identifying reliable literature one step further, we introduce here the concept of *High Utility Regions*. These are regions in the text that we intuitively characterize as focusing on scientific findings, stated with a high confidence, and preferably supported by experimental evidence. Following this line of reasoning, we devised criteria for characterizing statements made in the literature along several axes, which we describe further in the following sections. The axes that we introduce include *focus* (e.g. scientific vs. general), *polarity* (positive vs. negative statement), level of *certainty*, strength of *evidence*, and *direction/trend* (increase or decrease in certain measurement). The utility of a region, as a source for scientific knowledge, can be evaluated based on its "coordinates" along these axes. Earlier work on annotation of scientific text (e.g. [21,11,14]) focused on the partition of text into zones, according to the type of discourse and the components of scientific argumentation (e.g. *background, framework, aim*). In contrast, we define a set of dimensions, along which each statement in the text is characterized.

Our present study is motivated by the need to identify and characterize locations in published papers where reliable scientific facts can be typically found. We ultimately aim to develop machine-learning methods for classifying statements along these multiple dimensions.

While the above goals motivate ongoing work, underlying their realization is another major component, namely the establishment of a large and reliable body of annotated text for training and testing. In this manuscript we focus on several aspects of the corpus preparation while the learning tasks will be addressed elsewhere.

The planning and the building of the annotated corpus have two main beneficial outcomes: Composing a set of well-formed and tested guidelines for annotators, and generating an annotated corpus for testing and training text-mining methods. We believe that the lessons

learned, as well as the guidelines themselves, could be useful to other researchers.

The rest of the paper describes the work we have done to characterize phrases along the multiple axes mentioned above, and to build the training/test corpus around this characterization. We start with an overview of the dimensions used for characterizing text fragments, and follow with a more detailed description of the annotation guidelines. We then discuss a test we conducted to evaluate these guidelines by measuring inter-annotator agreement.

## Characterization Axes

We examine full-text scientific journal papers, from multiple domains of biomedical discourse. Each assertion in the corpus (where an *assertion* may be a sentence or just a fragment of a sentence, as described below) can be characterized by its type, and marked-up along the following dimensions, which we broadly define as follows:

- **Focus**. Indicates the type of the information conveyed by the assertion. Focus can be either:
  ***Scientific****:* Describes findings and discovery; Tagged by the letter *S.*
  ***Generic****:* General state of knowledge and science outside the scope of the paper, the structure of the paper itself, or the state of the world. Such statements are not usually based on scientific experiment, and would probably be as valid, if made by a layperson. Tagged as *G.*
  ***Methodology:*** Describes a procedure that was used to execute an experiment or a study, denoted by the letter *M.*
- **Polarity:** Indicates whether the assertion is made using *positive* terms (e.g. "we found that…", tagged as *P*) or *negative (*e.g. "There was no significant change in…", tagged as *N).*
- **Certainty:** Indicates the degree of certainty regarding the validity of the assertion. The annotation uses a scale of 0-3 to measure certainty level, of both positive and negative statements.
  The lowest degree (*0*) represents *complete uncertainty*, (e.g. "*it is unknown if*..." or "*it is unclear whether*..." etc.). The highest degree, (*3*), represents complete certainty (an accepted, known and/or proven fact). Intermediate *degrees:* (*1*) represents a low certainty, while (*2*) is assigned to *high-likelihood* expressions that are still short of complete certainty.
- **Evidence:** Indicates, for each fragment, if the assertion it makes is supported by experimental evidence. The evidence tag is the letter *E,* followed by a single digit, (0-3), indicating the type of evidence or its absence. The tag ***E0*** is used when there is no indication of evidence, as well as in cases where the text explicitly states *lack of evidence*. The tag ***E1*** indicates that a claim of evidence exists in the text. (e.g. "It was shown that…."), but the evidence itself is not given. The tag ***E2*** is used when the evidence is not given within the text, but explicit reference is made to another paper supporting the assertion. The tag ***E3*** is used when evidence is directly provided in the text, for instance, expressed as a reference to the experimental result reported within the paper (e.g. "our results show"…), a reference to a figure demonstrating the results, or other direct reference. (Further details are beyond the scope of this note).
- **Direction/Trend:** Indicates whether the assertion reports an *increase* or a *decrease* in a specific phenomenon, finding or activity. An increase is denoted by a "**+**" while a decrease is denoted by a "**–** ".

A significant advantage of separating the *polarity* from the *direction* is the provision of a straightforward way to handle occurrences of double-negation, and an almost completely mechanical way for annotators to tag such cases. The *polarity* refers to what the authors observed or did not observe. For instance "We have seen a significant…" has a *positive* (*P*) polarity while "There was no…" has a *negative* (*N*) polarity, regardless of what was observed. The state of the observed object is encoded in the *direction* or the *trend* indicated for the finding or observation. To continue our example, if the sentence is "We have seen a significant reduction in the expression level…" the sentence still has *positive polarity*, but its trend is negative (-). On the other hand the sentence "There was no significant increase in the expression of…" has a negative polarity, along with an upward trend ("increase) denoted by a +.

An important aspect of characterizing text using the tags above, is defining the appropriate units to which such tags are assigned. One could consider tagging several levels of text-units, from whole sections, through paragraphs, to sentences and individual phrases. Paragraphs usually contain too

much variation to merit a single tag. Individual sentences may vary greatly in scientific content and polarity. Moreover, even within sentences, there is often variability in content, polarity and the level of evidence.

Therefore we suggest that a separate tagging be assigned to each fragment within a sentence. Fragmentation takes place either when there is a change in content along any of the 5 dimensions listed above (e.g. a statement's polarity changes from positive to negative) or at conjunction points in compound sentences. Each tag starts with the ordinal number of the fragment within the sentence.

Figure 1 provides several examples of tags assigned to sentence fragments using the method discussed above. The first one is an unfragmented sentence, with a science focus, high confidence and experimental evidence indicated by the words "we demonstrate". The second example has a negative trend (-) due to the term "inhibited", while the third example has a negative polarity ("did not identify"). The 4th and 5th examples are of methodology and generic sentences, while the last example demonstrates the fragmentation of a sentence into three annotated fragments based on changes in polarity and direction.

We demonstrate that ICG-001 binds specifically to CBP. ***1SP3E3*

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

A recent Japanese study, for example, did not identify any exon 3 missense mutations. ***1SN3E1*

Anesthetized rats (methoxyflurane) were perfused with 4% buffered paraformaldehyde. ***1MP3E3*

Mechanistic arguments have been put forward for either pattern in humans. ***1G*

We demonstrate that ICG-001 binds specifically to CBP ***1SP2E3*
but not the related transcriptional coactivator p300, ***2SN2E3*
thereby disrupting the interaction of CBP with beta-catenin. ***3SP2E3-*

**Figure 1. Examples of annotation tags.**

## Testing the Guidelines

The guidelines evolved through numerous iterations in which they were used to tag fragments from articles, ranging in style from reviews to research publications, from several biomedical domains. We test the guidelines by applying them to a set of paragraphs extracted from several papers, and evaluating our own inter-annotator agreement. While our own familiarity with these guidelines may seem to bias the results, the annotators who will tag the corpus will undergo training which would bring them to a similar level of familiarity with these guidelines. The evaluation setting and the results are described below.

***Evaluation Procedure:*** As some of the text properties we examine are local (e.g. polarity) while others may depend on context (e.g. evidence), the evaluation corpus was built by first selecting whole paragraphs rather than individual sentences from scientific papers. The paragraphs were taken from 6 recent molecular biology articles (3 published in *Science* and 3 in *Cell*), representing the full diversity of the article sections (abstract, introduction, results, methods, and discussion), and covering a variety of styles including editorials, reviews, and research articles. This evaluation corpus comprises a total of 81 sentences.

Each of the 3 authors independently fragmented the individual sentences into "annotation units" and annotated each unit along the axes given above, using the annotation guidelines. We then measured the annotation agreement rate along each of the axes. We chose to use this measure rather than the well-known Kappa coefficient since there are clear shortcomings in using the common chance-based Null-model in the Kappa measure [22] when the three annotators are following common guidelines applied to the same corpus and are therefore clearly not unconditionally independent.

***Evaluation results:*** The average pair-wise inter-annotator agreements for individual text properties (axes) are shown in the Table 1.

| Axes | Focus | Pol. | Cert. | Evid. | Trend |
|---|---|---|---|---|---|
| Average Agreement | 0.83 | 0.81 | 0.70 | 0.73 | 0.81 |

**Table 1 Average Inter-Annotator Agreement**

Out of 81 sentences, 54 were fragmented at the same point by all three evaluators, resulting in a total of 62 fragments in which agreement can be easily measured. We also measured agreement rates for the rest of the 27 sentences by aligning their fragment annotations, but this analysis is beyond the scope of this paper. The average agreement rate between every pair of annotators, along each of the axes (Focus, Polarity, Certainty, Evidence and Trend) is summarized in Table 1. For the most part, these numbers reflect high rates of inter-annotator agreement. Detailed analysis (not shown) suggests that most of the annotation differences on *Polarity* occur when some annotators assign polarity to fragments with non-science focus while others do not. The relatively

low agreement rates on Evidence and Certainty suggest a need for revision of the annotation guidelines along these two axes before embarking on the annotation of the full-fledged corpus.

## Ongoing Work

This paper introduces the concept of high-utility regions in text, and discusses the first stages of work towards identifying such regions in the biomedical literature. Our contribution here is the definition of criteria towards identifying such text, and of annotation guidelines. A training and test corpus, which will encompass hundreds of full-text papers and thousands of annotated sentences is currently under construction, where multiple annotators will tag each piece of text. This part of the task will be underway when the paper is presented. We are working on the construction of appropriate classifiers to automatically assign such tags. This work will be the topic of future reports.

The potential value of this work is three pronged: In the construction of the corpus, in the definition of several significant axes for biomedical text classification, and in the characterization of high-utility regions – as measured by scientific content, evidence and certainty. Our work towards identifying such regions automatically is expected to affect both extraction and retrieval from the biomedical literature.

**References**


1. Baruch J. *Progress in programming for processing English language medical records.* Ann. N Y Acad Sci. 126(2):795-804. 1965.
2. Craven M, Kumlien J. *Constructing biological knowledge bases by extracting information from text sources.* Proc of ISMB. 77-86. 1999.
3. Cohen KB, Hunter L. *Natural Language Processing and Systems Biology.* In AI and Systems Biology, Springer Series on Comp. Biology. Dubitzky W. and Azuaje F. (Eds.). 2005.
4. de Bruijn B, Martin J. *Getting to the (c)ore of knowledge: mining biomedical literature.* Int J Med Inf. 67(1-3):7-18. 2002.
5. Eskin E, Agichtein E. *Combining Text Mining and Sequence Analysis to Discover Protein Functional Regions.* Proc. of the 9th Pacific Symposium on Biocomputing. 288-299. 2004.
6. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB: *A general natural-language text processor for clinical radiology.* J Am Med Inform Assoc., 1(2):161-174. 1994.
7. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES*: a natural-language processing system for the extraction of molecular pathways from journal articles.* Bioinformatics. 17 Suppl 1S74-S82. 2001.
8. Glenisson P. *et al. TXTGate : Profiling gene groups with text-based information*. Genome Biology. 5(6). 2004.
9. Hersh W, Bhupatiraju TR, Corley S. *Enhancing Access to the Bibliome: The TREC Genomics Track.* MedInfo 2004. 773-777. 2004.
10. Krauthammer M. *et al. Of truth and pathways: chasing bits of information through myriads of articles.* Bioinformatics. 18, Suppl 1S249-57. 2002.
11. Langer H, Lüngen H, Bayrel PS. *Text Type Structure and Logical Document Structure*. Proc. of the ACL Workshop on Discourse Annotation. 2004.
12. Leek TR. *Information Extraction Using Hidden Markov Models.* MSc thesis, Dept. of Computer Science, University of California, San Diego. 1997.
13. Light M, Qiu XY, Srinivasan P. *The Language of Bioscience: Facts, Speculations, and Statements In Between.* HLT-NAACL: BioLink'04. 17-24. 2004.
14. Mizuta Y, Collier N. *Zone Identification in Biology Articles as a Basis for Information Extraction*. Proc. of the JNLPBA. 2004.
15. Rindflesch TC, Aronson AR. *Ambiguity resolution while mapping free text to the UMLS Metathesaurus.* Proc. of the Annu Symp Comput Appl Med Care:240-244. 1994.
16. Shatkay H, Feldman R. *Mining the biomedical literature in the genomic era: an overview.* J Comput Biol, 10(6):821-855. 2003.
17. Shatkay H, Edwards S, Wilbur WJ, Boguski M. *Genes, Themes and Microarrays: Using Information Retrieval for Large Scale Gene Analysis*. Proc. of ISMB. 2000.
18. Swanson DR. *Online search for logically-related noninteractive medical literatures: A systematic trial-and-error strategy.* JASIS 40(5): 356-358. 1989.
19. Tanabe L, Smith LH, Lee JK, Scherf U, Hunter L, Weinstein, JN. *MedMiner: An internet tool for filtering and organizing biomedical information, with application to gene expression profiling.* BioTechniques. 27:1210-1217.1999.
20. Tanabe L, Wilbur WJ. *Tagging gene and protein names in biomedical text*. Bioinformatics. 18(8):1124-32. 2002.
21. Teufel S, Carletta J, Moens M. *An Annotation Scheme for Discourse-Level Argumantation in Research Articles*. Proc. of EACL, 1999.
22. Uebersax J. *Diversity of Decision-Making Models and the Measurement of Inter-rater Agreement.* Psychological Bulletin. 101(1): 140-146. 1987. Also see:*http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm*