Getting to the (C)ore of Knowledge: Mining Biomedical Literature

de Bruijn, Berry; Martin, Joel

International Journal of Medical Informatics, 67, 1-3, pp. 7-18, 2002-11-14

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

*Getting to the (C)ore of Knowledge: Mining Biomedical Literature**

De Bruijn, B., and Martin, J.
December 2002

* published in International Journal of Medical Informatics. Vol. 67 1-3, pp. 7-18, December 2002. NRC 45880.

Canada

National Research Council
Institute for Information Technology
1200 Montreal Road, building M50
Ottawa, ON, Canada K1A 0R6

NRC · CNRC

Getting to the (c)ore of knowledge: mining biomedical literature

Berry de Bruijn and Joel Martin
berry.debruijn@nrc.ca; joel.martin@nrc.ca

Abstract

Literature mining is the process of extracting and combining facts from scientific publications. In recent years, many computer programs have been designed to extract various molecular biology findings from Medline abstracts or full text articles. The present article describes the range of text mining techniques that have been applied to scientific documents. It divides 'automated reading' into four general subtasks: text categorization, named entity tagging, fact extraction, and collection-wide analysis. Literature mining offers powerful methods to support knowledge discovery and the construction of topic maps and ontologies. An overview is given of recent developments in medical language processing. Special attention is given to the domain particularities of molecular biology, and the emerging synergy between literature mining and molecular databases accessible through Internet.

Introduction

With an overwhelming amount of biomedical information available as text, it is natural to ask if it can be read automatically. For several decades, natural language processing (NLP) has been applied in biomedicine to automatically 'read' patient records and has resulted in a growing, but fairly homogeneous body of research. Now with the explosive growth of molecular biology research, there is a tremendous amount of text of a different sort, journal articles. The text collection in Medline can be mined to learn about a subfield, find supporting evidence for new experiments, add to molecular biology databases, or support Evidence Based Medicine.

Literature mining can be compared to reading and understanding literature but is performed automatically by a computer. Like reading, most literature mining projects target a specific goal. In bioinformatics, examples are:

- Finding protein-protein interactions [a.o. 1-3],
- Finding protein-gene interactions [4],
- Finding subcellular localization of proteins [5-7],
- Functional annotation of proteins [8, 9],
- Pathway discovery [10, 11],
- Vocabulary construction [12, 13, 14],
- Assisting BLAST or SCOP search with evidence found in literature [15, 16],
- Discovering gene functions and relations [17].

A few examples in medicine include:

- charting a literature by clustering articles [18]
- discovery of hidden relations between, for instance, diseases and medications [19-21]
- use medical text to support the construction of knowledge bases [22]

With this wide variety of goals, it is not surprising that many different tools have been adopted or invented by the various researchers. Although the approaches differ, they can all be seen as examples of one or more stages of a reading process.

Most of the studies that work with biomedical literature use Medline abstracts. This underlines the immense value of the Medline collection. Its size has passed the count of 12 million citations, most of which include abstracts. Our hope is that in future years, more and more initiatives will and can be directed towards the full text of articles. A number of publishers now offer free on-line access to full articles and standards in web lay-out and metatagging are finding their acceptance. Algorithms that scale up

This article was published as: Berry de Bruijn and Joel Martin: Getting to the (c)ore of knowledge: mining biomedical literature. International Journal of Medical Informatics 67 (2002) pp 7-18.

better and a continuous increase in affordable computing power are - or will be - ready to tackle that.

Free availability of material is at this moment trapped between two forces. There is the growing pressure from the (noncommercial) scientific community to freely share material. On the other end of the see-saw sits the growing pressure on companies to make a profit on the web and therefore to regulate access to material.
In biomedicine, the efforts of the U.S. National Library of Medicine are once more invaluable on this matter. PubMed Central aims to facilitate and/or host full-text access to participating journals in a common format, and requires that access is free at least one year after publication and preferably sooner than that. Currently, more than 25 journals have committed to this initiative.

This article reviews a number of studies on literature mining applied to biomedicine, and takes a look at the range of techniques that have been (or could be) applied to modules within the literature mining process. The nature of an article such as this, is that it can only present a snapshot of the state of the art at one point in time. For a more up-to-date overview of NLP studies applied to molecular biology and other biomedical domains see our on-line, partially annotated, extensive bibliography at http://textomy.iit.nrc.ca/cgi-bin/BNLPB_ix.cgi .

Very recently, an overview on Genomics and Natural Language Processing appeared [23]. That article is written from a genomics perspective, and as such concentrates partly on Information Retrieval techniques (possibly including a literature corpus) to support sequence finding and annotation. Our article is written from an NLP researchers' point of view, and reviews in what ways recent studies - notably in the area of molecular biology and literature searching - have added to the field of Natural Language Processing in Biomedicine. We see both articles complementing each other.

Natural language processing in biomedicine: a brief overview

The application of natural language processing for molecular biology might be relatively new, but NLP has been applied to biomedical text for decades, in fact, soon after computerized clinical record systems were introduced in the mid 1960s [24]. The computerization of clinical records increased the tension in the field of medical reporting and recording. Structured reporting, on the one hand, ensures rigidity and optimal retrievability of records. Natural language narrative, on the other hand, ensures flexibility and allows unequaled representation of detail and freedom of expression. Natural language processing techniques were adopted to pursue a 'best of both worlds' setting with as little tension as possible.

Spyns [25] wrote a broad overview of natural language processing in medicine, giving ample attention to milestone projects and systems such as the Linguistic String Project, Specialist, Recit, MedLEE, and Menelas. The overview of Friedman and Hripcsak [26] also concentrates on NLP with clinical narrative, giving a short summary of earlier projects and the state of the art at that point in time. We refer the reader to these studies for a complete overview, and concentrate in this section on newer articles, emerging trends, and developments that are directly relevant to the discussion in the remainder of this article.

In recent years, research has continued to focus on text indexing and document coding to allow powerful, meaningful retrieval of documents. Document indexing uses terms from a glossary or ontology (MeSH, UMLS, SNOMED) or text features such as words or phrases. The parameters of the feature selection algorithm can be used to tune a system towards higher precision, for instance by using multi-word phrases [27, 28], or better recall, for instance by using sub-word strings [29, 30]. Various methods have been applied to medical scientific literature [31-34] and to clinical narrative [35-37].

One major contrast between most NLP research in clinical medicine and the more recent ones in molecular biology is the type of language material: patient records vs. scientific articles. Most NLP systems in clinical medicine work with text from patient records such as discharge summaries and diagnosis reports. NLP systems in bioinformatics use mostly articles or abstracts from the scientific medical literature. Differences between these two types of text affect the choice of techniques for NLP. Biomedical literature is carefully constructed and meticulously proof-read, so spelling errors and incomplete parses are less of a problem. On the other hand, new concepts may be introduced, such as a newly unraveled molecule. The bulk of literature is in English.
Clinical narrative, on the other hand, might be more colloquial with the use of ungrammatical constructs and unstandardized abbreviations. It is more likely to

contain segments of 'canned text' - longer phrases or possibly entire paragraphs that are repeatedly encountered between records. Unknown words are, if not spelling errors, often proper names such as patient names, doctor names, addresses or institution names. Work on clinical narrative includes methods that handle other languages than English, or do cross language operations such as retrieval from multilingual collections, or interlingual translation [38-40].

In recent years, knowledge intense NLP methods have kept their ground and fortified it. Knowledge structures such as UMLS and Galen have grown and have become better accessible for NLP systems. Methodologies for using knowledge structures have become more refined [33, 34, 41]. Moreover, there is a circular amplification effect: better language processing systems help in the construction of better knowledge structures [22, 42].

At the same time, statistical methods for language processing have gained ground in recent years. Better availability of sizable corpora, combined with affordable hardware to store and process large amounts of text, have certainly fed this development. Examples of recent studies are: Chapman et al. [43] applied Bayesian networks for text categorization, Wilbur et al. [44] used a Bayes classifier for spotting chemical names in text; Taira et al. [45] used statistical machine learning methods to structure the contents of radiology reports. De Bruijn et al. [30, 35] introduced nearest neighbour classification techniques for assigning SNOMED terms to diagnostic narrative.

In recent years, many interesting and new applications for NLP have been presented. We mention a few of them here. Ruch et al. [46] used NLP to scrub patient names (and names of providers, institutions) from clinical reports so that anonymized reports can be used for research while observing the patient's privacy. Zweigenbaum and Grabar [47] showed that French texts that accidentally lost their accents, for instance after electronic processing, can be automatically re-accented with high accuracy despite various pitfalls. Liu et al. [48] applied unsupervised learning to disambiguate biomedical terms.

In the mean time, work on large scale projects has been going on, and new large scale projects have emerged. Hahn [22, 49] at Freiburg University describes MEDSYNDIKATE, the member of the SYNDIKATE family of knowledge-acquisition-from-text systems that is targeted towards medicine. The knowledge-rich infrastructure on which the SYNDIKATE systems rely are incrementally augmented from text and from external knowledge structures, such as UMLS.

The NLM Indexing Initiative (IND, see [31]) integrates a number of earlier approaches with the purpose of automatically proposing MeSH indexing terms for citation titles and abstracts. MetaMap [50] is part of IND. UMLS is used to bridge the gap between free text and the MeSH lexicon. Apart from lexical and semantic analysis of the text itself, additional candidate MeSH terms are found by searching for textually similar documents in Medline (k-Nearest Neighbour classification). Future work in the IND includes incorporation of new methods such as various machine learning techniques, word sense disambiguation, full text processing and subdiscipline indexing based on journal descriptors.

Text mining as a modular process

Text mining is a process very similar to reading. A reader first selects what they will read, then identifies important entities and relations between those entities, and finally combines this new information with other articles and other knowledge. This reading process (see figure 1) forms the backbone of this article. In the following sections, the various studies on text mining applied to molecular biology literature are aligned with this modular view of the reading process.

Document categorization

Document categorization, at its most basic, divides a collection of documents into disjoint subsets. This is also known as Document or Text *Classification*, but *categorization* is the most common term. The

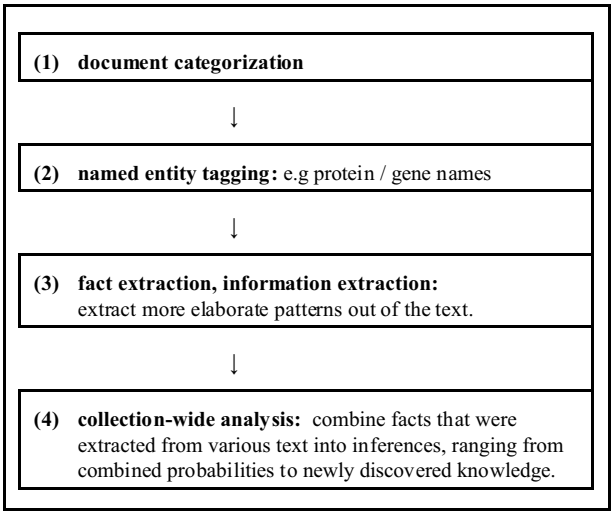

Figure 1: Text mining as a modular process.

categories are usually predefined; if they are not, the process is actually document clustering (grouping documents through their superficial characteristics, e.g., [51]). By this definition Information retrieval (IR) is one form of categorization: the collection is divided into two categories of documents, one relevant to the query and one irrelevant. IR algorithms however differ from more specialized categorization algorithms as they use queries rather than teaching from examples.

Document categorization is useful primarily for efficiency reasons. Automated readers, just like human readers, cannot usually spend the time to read all available documents. Having a relevant subset in an early phase can direct subsequent efforts, especially those that are computationally expensive. For example, a text mining system that hunts for subcellular localizations of proteins, might need one minute of processing time per Medline abstract. One can apply that system to all 12 million Medline abstracts and find in retrospect that only, say, 8,900 abstracts returned a valid finding. One could also use a document categorizer that finds, say, 10,000 promising abstracts, and see in retrospect that 8,800 abstracts were useful. A researcher might accept a slight loss of 100 documents with the huge reduction in processing time.

Document categorization can be used to aid human readers by providing a much more accurate, but slower and less flexible, alternative to search engines [e.g. 52]. Other projects explicitly include document categorization but as a module in a larger system [12, 53, 54]. Raychaudhuri [55] et al. used document categorization methods for finding - and so labeling - gene functions.

The methods used for document categorization can be borrowed from Machine Learning. Popular methods include Naive Bayes [52, 56], Decision Trees [57], Neural Networks, Nearest Neighbor [58] and Support Vector Machines [52, 59]. In all these methods, a collection of precategorized documents is used to train a statistical model of word or phrase use and then the statistical model is applied to uncategorized documents.

Before the training and the actual categorization, there are two preliminary steps: (1) feature extraction, and (2) feature set transformation. The characterizing features of documents can be based on words (most often), word combinations, character sequences or (more rarely) concepts associated with word occurrences. Feature set transformation has two purposes: reducing the size of the feature set, hoping that that will improve efficiency as well as effectiveness, and scaling or weighting the feature set with the purpose of improving the document representation relative to the entire collection. Reduction of the feature set is often done by stemming, eliminating stop words, and eliminating very rare words that burden the classifier more than that they add discrimination power. See for instance [52].

As one example, the Support Vector Machine (SVM) is a relatively new but promising technique for pattern categorization and it has been successfully applied to text [e.g. 59]. In an SVM, documents are represented as points in a vector space, where the dimensions are the selected features. Based on the training document vectors, the SVM finds the (unique) hyperplane that minimizes the expected generalization error. It does this by maximizing the shortest distance between any of the training examples and the hyperplane. Only some of the training vectors will finally define the position of the hyperplane so these are called the 'support vectors'. After the training phase, classification of new documents is a fast process. For biological literature, only few results have been reported. Wilbur [52] used an SVM in combination with a Naive Bayes classifier to construct a boosted system for text categorization. In our own project, we have been applying an SVM to various classes of abstracts and sentences from Medline with good results [60]. Advantages of SVM include its good and robust performance with typical accuracies of up to around 90% (precision=recall cut-off point), and resistance to overfitting the data.

The usual evaluation metric for document categorization tasks is accuracy (in multi-class systems), and the twin-metrics recall and precision (for binary class systems). It is often possible to tweak the system for better precision at the cost of recall or better recall at the cost of precision, so that a task-specific setting can be reached. In evaluation, this makes it possible to plot results in ROC curves. N-fold cross validation is the method of choice for evaluation.

Named entity tagging

The main reason to read an article is to find out what it says. Similarly, the goal of Information Extraction is to fill in a database record with specific information from the article. The first level of this task is to identify what entities or objects the article mentions. This is called named entity tagging, where

the beginning and end of entities might be marked with SGML or XML tags - see fig. 2.

In molecular biology, most of the entities are molecules, such as RNA, genes and proteins, and these entities have many aliases. The lack of naming conventions make this task more difficult. Molecule names are invented on a daily basis and conventions, if they exist, may differ between subdisciplines. Two molecules may share names, with only the context to distinguish between the gene and the protein. Even if names are not shared, a substring of an entity name might be a legitimate, but different entity. For example, tagging 'protein kinase 2' might be an adequate tag in a certain sentence, but 'protein kinase 2 alpha' might be even better. In medicine, an example of named entity tagging is identifying person names to ensure anonymity [46]. Such a 'scrubbing' system should be able to distinguish between 'Parkinson' as the name of a patient and as the reference to Parkinson's Disease.

All techniques suggested for finding named entities use some form of character-by-character or word-by-word pattern to identify the entities. In some of these techniques, the patterns are designed by hand. In others, the patterns are learned from examples that are provided by an expert. Then when a new article is encountered, each string of characters or words is scanned looking for close matches to the learned patterns.

The simplest, manual, approach is to take advantage of *string regularity* and write patterns to capture the known naming conventions, such as a 'p' preceding or succeeding a gene name [61]. Other reliable rules are possible that identify certain words with letters and digits.

A second approach is *lexicon based* that uses name lists to tag terms, or likely components of entity names [2, 62]. The success of this approach depends on the availability and the coverage of such lists, as well as on their stability over time.

A final manual approach is *context based*. In this method, a dictionary of sentence contexts is compiled that suggest likely molecule names. For instance, in a sentence that shows the pattern "<protein A> inhibits <unknown string>", a rule can dictate that the unknown string is a candidate protein name.
The learning methods, on the other hand, are applied when it is deemed impossible, inaccurate or too slow to manually compile the string regularities and lexicon and context dictionaries. Hishiki et al. [63] use a machine learning module to identify which sequences of n characters are likely to be a part of a molecule name. The most likely ones are the string regularities. New sequences are then scored by the system's past experience with such sequences.

Hidden Markov Models (HMMs) [64] can learn a lexicon and context as well by computing the probability that a sequence of specific words surround or constitute a molecule name. The expert just has to identify examples, while the HMM learns the patterns to apply to new sequences of words.

Above methods do not have to be used in isolation. Friedman et al. [10] used string regularity as well as a lexicon to tag protein and gene names. Also, the methods can be improved by filtering the text. Some researchers prefer to apply part-of-speech tagging to help the Named entity tagging task, so that only (whole) noun phrases are considered as candidate molecule names. The popular part-of-speech taggers or shallow parsers appear to be flexible enough to handle the specialized biological language. For instance, EngCG was used by Hishiki et al. [63] and by Yakushiji et al. [65]. Erikson et al. [66] combine an off-the-shelf syntactic parser (FDG) with hand written rules and a local dynamic dictionary.

For protein name tagging, accuracies as high as around 95% have been reported [67], but care should be given to the test set composition. It is known that for some organisms or some protein subdomains, the nomenclature is fairly rigidly standardized and excellent tagging accuracy can be reached there. Likewise, experiments with lower results should not be discarded without close scrutiny of the application domain: it might be that the study concentrates on a trickier problem.

'raw' sentence:	The interleukin-1 receptor (IL-1R) signaling pathway leads to nuclear factor kappa B (NF-kappaB) activation in mammals and is similar to the Toll pathway in Drosophila.
tagged sentence:	The <protein>**interleukin-1 receptor**</protein> (<protein>**IL-1R**</protein>) signaling pathway leads to <protein>**nuclear factor kappa B**</protein> (<protein>**NF-kappaB**</protein>) activation in mammals and is similar to the <protein>**Toll**</protein> pathway in <organism>*Drosophila*</organism>.

Figure 2: an example of named entity tagging on protein and organism names.

Fact extraction

Readers do not understand text if they merely know the entities. They must also grasp the interactions or relationships between those entities. Fact extraction is the identification of entities and their relations. To have a machine do this correctly for arbitrary relationships would require a full natural language intelligence, something that is many years away. There are several approximations that have been tried, from purely statistical co-occurrence to imperfect parsing and coreference resolution.

The simplest approach to capture entity relationships is to search for sentences that mention two entities of the right sort frequently enough. For example, the frequent co-occurrence of two specific protein names with a verb that indicates a molecular interaction might be enough to guess the existence of such an interaction. Craven [5] had his system find sentences where a protein name occurred together with a subcellular location. The effect of accidental co-occurrence could be minimized by requiring frequent corroboration of any pairing. Using a similar co-occurrence approach, Ding et al [67] found that precision and recall traded off when the length of the used text segment was varied. Working with phrases gave generally better precision, while working with entire abstracts gave best recall; sentences scored in between. Jenssen et al. [68] searched for gene name co-occurrence in abstracts and then added more meaning to the relation by annotating it with selected MeSH terms from the document.

Another approach that increases the reliability of discovered relationships searches for fixed regular linguistic templates [2, 3, 69]. For example, the system might search for a specific interaction verb while verifying that the surrounding context is parsable in a correct syntactic structure and with entity names in the allocated positions - taking any (negative) modifiers into account - and only then assume the interaction between the substances to be sufficiently proven. The main disadvantage of this approach is that usually the templates must be constructed by hand. Also, many relationships that do not match the template will be missed, but a few good patterns (even when they have low recall) might extract a good number of facts out of a large corpus.

Some linguistic templates can be learned, for instance using a Hidden Markov Model [70]. This requires a corpus with annotated patterns - something that is harder to find or more labour-intensive to construct than a named entity annotated corpus. The expert must mark both the entities and which of several relations applies between those entities. There are clear advantages, no need to explicitly craft rules, better 'portability', and possibly greater overall recall.

Finally, even though automated understanding is not fully possible, important relationships can be discovered by performing a full syntactic parse, where relations between syntactic components are inferred [65, 71, 72]. This approach is similar to the template searching except that it is not domain specific and attempts to identify many or all relationships in a sentence. Park [73] illustrates the syntactical complexities and pitfalls of sentences in biomedical documents.

As an alternative to developing a literature mining system from scratch, some groups have adapted systems or modules of earlier developed systems. They were originally conceived for other bioinformatics tasks (Jake, Kleisli [11, 54]), for other medical domains (e.g. MedLEE [10], MedSynDiKaTe [49]) or for general use (e.g. Highlight [3], LaSIE [74]).

Collection wide analysis

Thinking new thoughts and using what is known, requires integrating information between documents. This opens the door to knowledge discovery, where combined facts form the basis of a novel insight. The well-known Swanson study [19, 20] on the relation between Raynauds disease and fish oil, was a starting point of formal literature-based knowledge discovery. Weeber et al. [21] discuss an automated replication of that study and similar ones.

Other studies have addressed knowledge discovery in molecular biology (see [5, 10]). As an example: from document 1 you were able to extract the relation “A implies B”; from document 2 you deduced that “B implies C”. So you might want to study whether “A implies C”, for which you have found no previous evidence in the literature.

Blaschke et al. [1] used a large number of automatically extracted facts on protein-protein interactions to graph an interaction map for the drosophila cell cycle. This is one illustration where the system abstracts many articles and leaves it to the researcher to make inferences based on the output graph. Leroy and Chen [75] sketch the architecture of GeneScene, a system under development that aims to assist researchers in reviewing large numbers of articles by extracting information and thus

interconnecting the literature. Krauthammer et al. [76] use bibliometrical methods to portray the development of research ideas in literature, as well as to expand a biological knowledge base. A NLP module in the core of their system collects the elementary statements from publications. Wilbur [18] discusses the possibilities of automatically carving up the literature on a certain subject - his example is the literature on AIDS with over 50,000 Medline abstracts - into coherent clusters. Every cluster is then labeled with a thematic summary. Such algorithms can improve access to a document collection and improve human comprehension of a subject. Srinivasan [77] introduces tools to explore the literature through contingency tables on MeSH term co-occurrence. Although the study uses MeSH terms rather than free text, it ties in well with the other studies on text mining.

Less ambitious goals have still benefitted from collection-wide analyses. One notable application is using collection redundancy to compensate for recall limitations of both statistical and structural methods (e.g. [5, 68]). A high precision/fair recall algorithm such as the typical structural one should have a pretty good confidence in any fact that did get extracted. Facts that were missed in one document might get extracted from another if the fact is redundant. If higher recall with fair precision algorithm is achieved - something that statistical methods tend to do - the combined confidence from various redundant instances might be enough to accept an extracted fact (e.g. [8]).

Apart from findings from other documents in the collection, external sources might help the text analysis. Analogous to clinical settings where medical thesauri and classification schemes (MeSH, ICD, Snomed, ULMS) are used to support text algorithms, database structures in biology (such as GenBank, SwissProt) can be applied towards the correct analysis of abstracts or full text. Craven [5] used Yeast Protein Database data, Krauthammer [62] used BLAST for protein name tagging; Hatzivassiloglou [78] mentions validation across other publications and existing knowledge. Such an integration of knowledge and cross referencing of literature can ultimately lead to a tight ontological structure [34, 79]. Yandell and Mandoros [23] see the synergistic tie between language processing systems and ontologies as a main promise for the future.

With higher hopes on collection-wide analyses, the scalability of algorithms becomes a more urgent issue. Considering the current size of Medline (close to 12 million articles) and its growth rate, and considering that full text articles are getting more and more available in electronic form, practical algorithms should scale up well. The ever-increasing power of computers helps in that respect too.

Concluding remarks

This overview showed a very wide variety of current applications and techniques for literature mining on biomedical text. The field is likely to become only wider in the future. On-line access of molecular databases and medical knowledge structures will augment the knowledge component in literature mining systems. Large-scale statistical methods will continue to challenge the position of the more syntax-semantics oriented approaches, although both will hold their own place. Literature mining systems will move closer towards the human reader, supporting subtasks of reading in a more interactive and flexible way. For instance by doing text categorization and named entity tagging on-the-fly, working with training material that can easily be edited and augmented.

Written language will always remain only semi-structured - and we see that as a benefit. Literature mining adds to written language the promise of making translations onto structures that we do not yet foresee. Therefore, these methods will continue to be fruitful even when some of the molecular biology challenges are solved.

References

[1] C. Blaschke, M.A. Andrade, C. Ouzounis, Valencia A: Automatic extraction of biological information from scientific text: protein-protein interactions. Proc Int Conf Intell Syst Mol Biol 1999;30A(2):60-7.

[2] T. Ono, H. Hishigaki, A. Tanigami, T. Takagi: Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics 2001 Feb;17(2):155-61.

[3] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, M. Carroll: Automatic extraction of protein interactions from scientific abstracts. Pac Symp Biocomput 2000:541-52.

[4] T. Sekimizu, H.S. Park, J. Tsujii: Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts. Genome Inform Ser Workshop Genome Inform 1998;9:62-71.

[5] M. Craven: Learning to Extract Relations from MEDLINE. AAAI-99 Workshop on Machine Learning for Information Extraction - July 19, 1999, Orlando Florida

[6] M. Craven, J. Kumlien: Constructing biological knowledge bases by extracting information from text sources. Proc Int Conf Intell Syst Mol Biol 1999:77-86.

[7] B.J. Stapley, L.A. Kelley, M.J. Sternberg: Predicting the sub-cellular location of proteins from text using support vector machines. Pac Symp Biocomput 2002;:374-85.

[8] M.A. Andrade, A. Valencia: Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. Proc Int Conf Intell Syst Mol Biol 1997;5(2):25-32.

[9] A. Renner, A. Aszodi: High-throughput functional annotation of novel gene products using document clustering. Pac Symp Biocomput 2000:54-68.

[10] C. Friedman, P. Kra, H. Yu, M. Krauthammer, A. Rzhetsky: GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 2001 Jun;17 Suppl 1:S74-S82.

[11] S.K. Ng, M. Wong: Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts. Genome Inform Ser Workshop Genome Inform 1999;10:104-112.

[12] Y. Ohta, Y. Yamamoto, T. Okazaki, I. Uchiyama, T. Takagi: Automatic construction of knowledge base from biological papers. Proc Int Conf Intell Syst Mol Biol 1997;5(2):218-25.

[13] T.C. Rindflesch, L. Hunter, A.R. Aronson: Mining molecular binding terminology from biomedical text. Proc AMIA Symp 1999;34(12):127-31.

[14] M. Yoshida, K. Fukuda, T. Takagi: PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary. Bioinformatics 2000 Feb;16(2):169-75.

[15] J.T. Chang, S. Raychaudhuri, R.B. Altman: Including biological literature improves homology search. Pac Symp Biocomput 2001;24(1):374-83.

[16] R.M. MacCallum, L.A. Kelley, M.J. Sternberg: SAWTED: structure assignment with text description--enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. Bioinformatics 2000 Feb;16(2):125-9.

[17] H. Shatkay, S. Edwards, W.J. Wilbur, M. Boguski: Genes, themes and microarrays: using information retrieval for large-scale gene analysis. Proc Int Conf Intell Syst Mol Biol 2000;8:317-28.

[18] W.J. Wilbur: A thematic analysis of the AIDS literature. Pac Symp Biocomput 2002;:386-97.

[19] D.R. Swanson.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 1986 Autumn; 30(1):7-18

[20] D.R. Swanson.: Medical literature as a potential source of new knowledge. Bull Med Libr Assoc 1990 Jan;78(1):29-37

[21] M. Weeber, H. Klein, A.R. Aronson, J.G. Mork, L.T. de Jong-van den Berg, R. Vos: Text-based discovery in biomedicine: the architecture of the DAD-system. Proc AMIA Symp 2000;35(20): 903-7.

[22] U. Hahn, M. Romacker, S. Schulz: Automatic Knowledge Engineering in Medicine: The MEDSYNDIKATE Text Mining System. NLPBA 2002 - EFMI Workshop on Natural Language Processing in Biomedical Applications, Nicosia Cyprus 2002: 41-45

[23] M.D. Yandell, W.H. Majoros: Genomics and natural language processing. Nat Rev Genet 2002 Aug;3(8):601-10.

[24] J.J. Baruch: Progress in programming for processing English language medical records. Ann N Y Acad Sci 1965 Aug 6;126(2):795-804.

[25] P. Spyns: Natural language processing in medicine: an overview. Methods Inf Med 1996 Dec;35(4-5):285-301.

[26] C. Friedman, G. Hripcsak: Natural language processing and its future in medicine. Acad Med 1999 Aug;74(8):890-5.

[27] D.C. Berrios: Automated indexing for full text information retrieval. Proc AMIA Symp 2000;:71-5.

[28] Baud RH, Lovis C, Rassinoux AM, Scherrer JR: Alternative ways for knowledge collection, indexing and robust language retrieval. Methods Inf Med 1998 Nov;37(4-5):315-26.

[29] U. Hahn, M. Honeck, M. Piotrowski, S. Schulz: Subword segmentation--leveling out morphological variations for medical document retrieval. Proc AMIA Symp 2001;:229-33.

[30] L.M. de Bruijn, A. Hasman, J.W. Arends: Supporting the classification of pathology reports: comparing two information retrieval methods. Comput Methods Programs Biomed 2000 Jun;62(2):109-13.

[31] A.R. Aronson, O. Bodenreider, H.F. Chang, S.M. Humphrey, J.G. Mork, S.J. Nelson, T.C. Rindflesch, W.J. Wilbur: The NLM Indexing Initiative. Proc AMIA Symp 2000;:17-21.

[32] K. Baclawski, J. Cigna, M.M. Kokar, P. Mager, B. Indurkhya: Knowledge representation and indexing using the unified medical language system. Pac Symp Biocomput 2000;:493-504.

[33] O. Bodenreider: Using UMLS semantics for classification purposes. Proc AMIA Symp 2000;28(10):86-90.

[34] B. Jackson and W. Ceusters: A novel approach to semantic indexing combining ontology-based semantic weights and in-document concept co-occurrences. NLPBA 2002 - EFMI Workshop on Natural Language Processing in Biomedical Applications, Nicosia Cyprus 2002:75-80

[35] L.M. de Bruijn, A. Hasman, J.W. Arends: Automatic coding of diagnostic reports. Methods Inf Med 1998 Sep;37(3):260-5.

[36] P. Franz, A. Zaiss, S. Schulz, U. Hahn, R. Klar: Automated coding of diagnoses--three methods compared. Proc AMIA Symp 2000;76(4):250-4.

[37] R.C. Barrows Jr, M. Busuioc, C. Friedman: Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. Proc AMIA Symp 2000;7(6):51-5.

[38] S. Schulz, U. Hahn: Morpheme-based, cross-lingual indexing for medical document retrieval. Int J Med Inf 2000 Sep;58-59:87-99.

[39] J.M. Rodrigues, B. Trombert-Paviot, R. Baud, J. Wagner, F. Meusnier-Carriot: Galen-In-Use: using artificial intelligence terminology tools to improve the linguistic coherence of a national coding system for surgical procedures. Medinfo 1998;9 Pt 1:623-7.

[40] Baud R, Lovis C, Rassinoux AM, Michel PA, Scherrer JR: Automatic extraction of linguistic knowledge from an international classification. Medinfo 1998;9 Pt 1:581-5.

[41] M.B. do Amaral, A. Roberts, A.L. Rector: NLP techniques associated with the OpenGALEN ontology for semi-automatic textual extraction of medical knowledge: abstracting and mapping equivalent linguistic and logical constructs. Proc AMIA Symp 2000;164(11):76-80.

[42] C. Friedman: A broad-coverage natural language processing system. Proc AMIA Symp 2000;19(19):270-4.

[43] W.W. Chapman, M. Fizman, B.E. Chapman, P.J. Haug: A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. J Biomed Inform 2001 Feb;34(1):4-14.

[44] W.J. Wilbur, G.F. Hazard Jr, G. Divita, J.G. Mork, A.R. Aronson, A.C. Browne: Analysis of biomedical text for chemical names: a comparison of three methods. Proc AMIA Symp 1999;:176-80.

[45] R.K. Taira, S.G. Soderland, R.M. Jakobovits: Automatic Structuring of Radiology Free-Text Reports. Radiographics

2001 Jan;21(1):237-245.
[46] P. Ruch, R.H. Baud, A.M. Rassinoux, P. Bouillon, G. Robert: Medical document anonymization with a semantic lexicon. Proc AMIA Symp 2000;:729-33.
[47] P. Zweigenbaum, N. Grabar: Accenting unknown words: application to the French version of the MeSH. NLPBA 2002 - EFMI Workshop on Natural Language Processing in Biomedical Applications, Nicosia Cyprus 2002: 69-74.
[48] H. Liu, Y.A. Lussier, C. Friedman: Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. abstract -- J Biomed Inform 2001 Aug;34(4):249-61.
[49] U. Hahn, M. Romacker, S. Schulz: Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. Pac Symp Biocomput 2002;:338-49.
[50] A.R. Aronson: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001;:17-21.
[51] I. Iliopoulos, A.J. Enright, C.A. Ouzounis: Textquest: document clustering of Medline abstracts for concept discovery in molecular biology. Pac Symp Biocomput 2001:384-95.
[52] W.J. Wilbur: Boosting naive Bayesian learning on a large subset of MEDLINE. Proc AMIA Symp 2000:918-22.
[53] L. Tanabe, U. Scherf, L.H. Smith, J.K. Lee, L. Hunter, J.N., Weinstein: MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques 1999 Dec; 27(6): 1210-1217.
[54] L. Wong: PIES, a protein interaction extraction system. Pac Symp Biocomput 2001:520-31.
[55] S. Raychaudhuri, J.T. Chang, P.D. Sutphin, R.B. Altman: Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. Genome Res 2002 Jan;12(1):203-14.
[56] E.M. Marcotte, I. Xenarios, D. Eisenberg: Mining literature for protein-protein interactions. Bioinformatics 2001 Apr;17(4):359-363.
[57] A. Wilcox, G. Hripcsak: Classification algorithms applied to narrative reports. Proc AMIA Symp 1999;20(1-2):455-9.
[58] L.M. de Bruijn, A. Hasman, J.W. Arends: Automatic SNOMED classification--a corpus-based method. Comput Methods Programs Biomed 1997 Sep;54(1-2):115-22.
[59] S. Dumais: Using SVMs for text categorization. IEEE Intelligent Systems 13(4), 1998 pp 21-23.
[60] B. de Bruijn, J. Martin, C. Wolting, I. Donaldson: Extracting sentences to justify categorization. ASIST Annual Meeting, Washington DC, Nov 2001: 460-457.
[61] K. Fukuda, A. Tamura, T. Tsunoda, T. Takagi: Toward information extraction: identifying protein names from biological papers. Pac Symp Biocomput 1998;33(2):707-18.
[62] M. Krauthammer, A. Rzhetsky, P. Morozov, C. Friedman: Using BLAST for identifying gene and protein names in journal articles. Gene 2000 Dec 23;259(1-2):245-52.
[63] T. Hishiki, N. Collier, C. Nobata, T. Okazaki-Ohta, N. Ogata, T. Sekimizu, R. Steiner, H.S. Park, J. Tsujii: Developing NLP Tools for Genome Informatics: An Information Extraction Perspective. Genome Inform Ser Workshop Genome Inform 1998;9:81-90.
[64] N. Collier, C. Nobata and J. Tsujii: Extracting the names of genes and gene products with a Hidden Markov Model.COLING 2000 conference proceedings, pp. 201-207
[65] A. Yakushiji, Y. Tateisi, Y. Miyao, J. Tsujii: Event extraction from biomedical papers using a full parser. Pac Symp Biocomput 2001:408-19.
[66] G. Eriksson, K. Franzen, F. Olsson, L. Asker and P. Linden: Exploiting Syntax when detecting Protein Names in Text. NLPBA 2002 - EFMI Workshop on Natural Language Processing in Biomedical Applications, Nicosia Cyprus 2002: 29-34.
[67] J. Ding, D. Berleant, D. Nettleton, E. Wurtele: Mining MEDLINE: abstracts, sentences, or phrases? Pac Symp Biocomput 2002;:326-37.
[68] T.K. Jenssen, A. Laegreid, J. Komorowski, E. Hovig: A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 2001 May;28(1):21-8.
[69] F. Eisenhaber, P. Bork: Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. Bioinformatics 1999 Jul-Aug;15(7-8):528-35.
[70] S. Ray, M. Craven: Representing Sentence Structure in Hidden Markov Models for Information Extraction. IJCAI 2001:1273-1279.
[71] T.C. Rindflesch, L. Tanabe, J.N. Weinstein, L. Hunter: EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput 2000:517-28.
[72] J. Pustejovsky, J. Castano, J. Zhang, M. Kotecki, B. Cochran: Robust relational parsing over biomedical literature: extracting inhibit relations. Pac Symp Biocomput 2002;:362-73.
[73] J.C. Park: Using combinatory categorial grammar to extract biomedical information. IEEE Intelligent Systems16(6):62-67.
[74] K. Humphreys, G. Demetriou, R. Gaizauskas: Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. Pac Symp Biocomput 2000;12(4):505-16.
[75] G. Leroy, H. Chen: Filling preposition-based templates to capture information from medical abstracts. Pac Symp Biocomput 2002;:350-61.
[76] M. Krauthammer, P. Kra, I. Iossifov, S.M. Gomez, G. Hripcsak, V. Hatzivassiloglou, C. Friedman, A. Rzhetsky: Of truth and pathways: chasing bits of information through myriads of articles. Bioinformatics 2002 Jul;18 Suppl 1:S249-S257.
[77] P. Srinivasan: MeSHmap: a text mining tool for MEDLINE. Proc AMIA Symp 2001;:642-6.
[78] V. Hatzivassiloglou, P.A. Duboue, A. Rzhetsky: Disambiguating proteins, genes, and RNA in text: a machine learning approach. Bioinformatics 2001 Jun;17 Suppl 1:S97-S106.
[79] H. Mima, S. Ananiadou, G. Nenadic and J. Tsujii: TIMS - A Workbench for Ontology-based Knowledge Acquisition and Integration. NLPBA 2002 - EFMI Workshop on Natural Language Processing in Biomedical Applications, Nicosia Cyprus 2002: 11-15